The Dissertation Committee for Min Gui Certifies that this is the approved version of the following dissertation:

Investigating the construct validity of the reading comprehension section of the College English Test in China: A structural equation modeling approach

Committee:

Elaine K. Horwitz, Supervisor

Diana C. Pulido

Min Liu

Thomas J. Garza

Tiffany A. Whittaker

Investigating the construct validity of the reading comprehension section of the College English Test in China: A structural equation modeling approach

by

Min Gui, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2011

Acknowledgements

I would like to express my sincere appreciation to all the people who supported me throughout the process of writing this dissertation. My supervisor, Dr. Elaine Horwitz, guided me through the challenges of this project. Her influence on me goes far beyond the dissertation. She is one of the two people who have enhanced the trajectory of my entire life. I have benefited tremendously from her wisdom, knowledge, and personality.

My committee members improved this study. Dr. Diana Pulido, who spent numerous hours with me, provided detailed guidance with respect to the available literature on reading research and the measurement of reading components skills. Dr. Tiffany Whittaker was always accessible. She responded to my emails within seconds and opened her office to me within the hour. Her support strengthened the research methodology of this project. Furthermore, she never forgot to boost my morale during the long journey of this study. I am thankful to Dr. Thomas Garza for his suggestions about my proposal. His encouragement and confidence in me were invaluable. Dr. Min Liu was kind in joining the committee, and she was always flexible and willing to help.

My dissertation benefited from the knowledge of Dr. Lia Plakans and Dr. William Grabe, although they were not on my committee. Dr. Lia Plakans, who motivated my interest in language assessment, was always there giving feedback even after she left the University of Texas at Austin. Dr. William Grabe's influence on the dissertation consisted of his informative guidance in the literature.

Many thanks go to Mengxuan Bai and Jing Chen for their assistance in administering the test and scoring. I appreciate their patience, time, and effort. I also need to thank the students who participated in the study.

I would like to extend my appreciation to my best friends in Austin, Chen Li and Lisal MacDonald. Their family-like support relieved my worries and increased my sense of security during my overseas study.

Finally, I am grateful to my husband and my daughter for their sacrifice, unwavering support, and unconditional love.

Investigating the construct validity of the reading comprehension section of the College English Test in China: A structural equation modeling approach

Min Gui, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Elaine K. Horwitz

The College English Test (CET) in China is the largest language test in the world. The number of CET test-takers has steadily increased from 100,000 for its first administration in 1987 to 13 million in 2006. CET scores are used to draw inferences about the test-takers' English as a foreign language proficiency as well as their specific skills in listening, speaking, reading, and writing. To justify the inferences drawn from test scores, evidence from a variety of sources should be constantly collected (Cronbach & Meehl, 1955; Messick, 1992; Chapelle, 1998; Bachman, 2000; Weir, 2005).

Despite the large-scale and high-stakes nature of the CET and the importance of test validation, studies on the quality of the CET are scarce. This study aims to examine the construct validity of the reading comprehension section of the CET by modeling the internal relationships between test-takers' scores on the CET reading section and their underlying reading abilities. Six components have been chosen as observed variables of

the latent variable of reading ability, namely, word recognition efficiency, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive reading skills. A pseudowords identification task programmed by the DMDX computer software, a revised version of Daneman & Carpenter's (1980) sentence reading span working memory test, Meara & Milton's (2002) Yes/No vocabulary tests, the syntactic test used in Shiotsu & Weir's (2007) study, Abeywickrama's (2007) discourse knowledge test, and a revised version of Phakiti's (2008) strategy use questionnaire were utilized to measure these six observed variables.

A total of 181 Chinese undergraduates participated in the study. With a baseline confirmatory factor model of reading ability and the CET scores, a structural model was analyzed. The results indicated that the path from reading ability to test performance was .75 and the squared regression coefficient of test performance was .56, which implied that participants' test performance was strongly underlined by their actual reading ability. Therefore, the scores on the CET reading section are largely justifiable for use in drawing inferences about participants' reading ability. Implications for validation research and reading instruction were also explored.

Table of Contents

List of Tables

List of figures

Chapter 1 Introduction

The purpose of this study is to investigate the construct validity of the reading comprehension section of the College English Test (CET) in China by examining the internal relationships between test-takers' CET reading scores and their underlying reading abilities. Confirmatory factor analysis (CFA) was employed to estimate reading ability. Six theoretically and empirically proposed components: word recognition efficiency, working memory, vocabulary knowledge, syntactic knowledge, discourse knowledge, and metacognitive reading strategies are used as the observed variables. The present study examined the extent to which test-takers' performances on the CET reading section is attributed to their reading ability. The results may serve to support or question the validity of the scores of the CET reading section, a major high-stakes test in China.

1.1 BACKGROUND

This study aims to investigate the construct validity of the reading comprehension section of the College English Test (CET) in China by analyzing the relationship between test scores and test-takers' underlying reading abilities through the use of structural equation modeling. Validity and fairness are the key issues that testing professionals should consider when evaluating a test (Bachman, 2000; Kunnnan, 1998; 2000), and, of course, without validity, there can be no fairness. Evidence from a variety of sources should be regularly gathered in order to justify the inferences drawn from test scores (Weir, 2005; Chapelle et al. 2008; Bachman, 2000; Chapelle, 1998; Messick, 1992). This

ongoing process of evidence collection is called validation. McNamara (2007) commented that "A language test without validation research is like a police force without a court system, unfair and dangerous" (p.280).

Despite the enormous importance of validation, studies on the validity of the CET are scarce. The paucity of research on the validity of the CET is particularly surprising in light of the fact that it is the largest English language test in the world with about nine million test-takers each year, a number approximately ten times larger than the TOEFL. After Yang & Weir's (1998) validation study on the CET only a handful of subsequent studies have been conducted. Furthermore, since Yang & Weir's study, the CET has experienced three important phases of change. The original CET test was composed of five sections: listening comprehension, reading comprehension, vocabulary and structures, cloze, and writing. In the first change, compound dictation and translation tasks were incorporated into the test from 1996. Second, the Spoken English Test (SET) was incorporated in 1999 as a component of the CET, although it is administered separately from the written test. Finally, the CET experienced a dramatic transformation in 2005: the listening section, which originally comprised 20% of the total score, was increased to 35%; the vocabulary and structure section were excluded; and writing was moved from the final testing task to the second, immediately following the fast reading section. In the reading section, the original four careful reading passages with 20 total questions were decreased to two careful reading passages with a total of 10 questions. A

fast reading (10%) and a cloze with a word bank (5%) were introduced at this time as a part of the reading section.

In sum, significant changes have been made to the testing tasks and the relative weightings of listening, reading, and writing tasks in the CET over the past 24 years (1987−2011). However, few studies have been conducted to justify these changes.

Serious problems have arisen with respect to the use of the CET and its scores, particularly as they apply to College English learning and teaching in China. First of all, having observed negative effects of the CET, a number of applied linguistics scholars have challenged the validity as well as the usefulness of the CET (e.g., Han, 2002; Liu, 2003; Qian, 2003; Han et al., 2004; Wang, 2006). Some scholars (e.g., Liu, 2003; Han et al., 2004) even posit, perhaps tongue in cheek, that it is high time the CET retire after having rendered meritorious service. Without sound and appropriate validation studies, the very existence of the CET is in question. Second, lacking information about the reasons for the change of test tasks, students tend to focus on practice of the new test forms. Ideally, any changes in the test should be based on research results and should result in a stronger internal linkage between test scores and actual language proficiency. Third, it would seem to be difficult for teachers to stress the importance of college English courses without a documented linkage between course performance and CET scores. Consequently, some teachers in China have become increasingly tolerant of students' absences from class and even accommodate student requests to cover test preparation materials in class rather than prioritizing the development of language skills.

The present study will focus on the reading section of the CET, which is one of the two largest components of the exam (the other is listening comprehension) and the internal relationship between CET reading scores and actual reading ability. In the following sections, the CET and the construct of reading ability and reading assessment will be briefly introduced, as will the theoretical background of the construct of reading ability and reading assessment. Lastly, the research questions will be presented.

1.2 A BRIEF INTRODUCTION TO THE CET

The CET is a paper and pencil test battery that is comprised of the College English Test Band 4 (CET-4), the College English Test Band 6 (CET-6), and the Spoken English Test (CET-SET). It is administered by the National College English Testing Committee on behalf of the Higher Education Department of the Ministry of Education of China. The CET-4 and the CET-6 are administered twice per year, in June and December. Both tests are 125 minutes in length. The CET-SET is also held twice per year, one month prior to the CET-4 and the CET-6. The CET-SET lasts about 20 minutes for a group of three test candidates. The first CET test was administered in 1987 with 100,000 test-takers (Zheng & Cheng, 2008). In the following years, the number has greatly increased. In 2003 9.15 million students participated in the test and in 2006 the number soared to 13 million ((Jin & Yang, 2006; Zheng & Cheng, 2008). The CET test battery purports to measure Chinese undergraduates' English language proficiency and to promote the implementation of the National College English Teaching Syllabus, as well

as to enhance the quality of teaching and learning of college English (Yang & Weir, 1998; Yang, 2000).

The present study will focus on the CET-4, which is taken prior to other two tests. The CET-4 is administered to all undergraduates, the CET-6 is offered to those who have passed the CET-4, and the CET-SET is directed for students with high CET-4 scores (550 and above) and those who obtain 520 and above on the CET-6. Furthermore, the CET-4 has the highest stakes of the three tests. Some universities set certain CET-4 scores as a requirement for graduation.

The CET reports four subscores as well as the total score: listening comprehension (249 points, 35%); reading comprehension (249 points, 35%); cloze or error correction (70 points, 10%); and writing and translation (142 points, 20%). The highest possible total score is therefore 710.

The reading test tasks, the focus of the current study, typically consist of

- 10 multiple-choice items based on a fast reading passage of approximately 1, 000 words;
- 10 multiple-choice questions based on two passages of 300−350 words each;
- 10 cloze blanks in a passage of 300−350 words to be filled in with one of the 15 words provided or 10 short answer questions based on a passage of a similar length.

These test items are intended to measure Chinese undergraduates' reading ability in English as a foreign language. The research questions of the present study center on whether the CET reading tasks measure the construct of reading ability, so a critical step is to examine how psychologists and reading researchers conceptualize the construct of reading and reading ability.

1.3 CONCEPTUALIZING READING AND READING ABILITY

A crucial procedure of the present study deals with modeling second language (L2) reading ability. L2 reading research has been tremendously influenced by first language reading research (Bernhardt, 2000; Grabe & Stoller, 2002; Grabe, 2009; Koda, 2005). Therefore, to elucidate the construct of L2 reading comprehension and present the rationale for the approach used to estimate L2 reading ability in the present study, it is necessary to review the research on L1 reading in general and the ways in which L1 reading comprehension and ability are conceptualized.

1.3.1 First language reading research

Chronologically, research on the perceptual processes of reading might roughly be divided into three periods. The first period lasts from the late 1870s to the early 1920s. The second phase of this research dates from the 1920s to the 1960s. The third era begins in the 1960s and continues to the present day.

The field of L1 reading research started with the establishment of Wundt's laboratory in Leipzig, Germany, in 1879 (Venezky, 1984; Rayner & Pollatsek, 1989).

The publication of Huey's (1908) work, *The Psychology and Pedagogy of Reading,* is the culmination of the research on reading of the first period. Huey and his contemporaries, who included Cattell, Thorndike, and Javal, focused their research on eye movements, perceptual span, word recognition, and reading rates.

The period from the 1920s to the 1960s was dominated by behaviorism, the doctrine of which holds that research should involve activities that can be observed. Since cognitive processes in reading are largely unobservable, the research on basic reading processes was not galvanized during the second period (Venezky, 1984; Rayner & Pollatsek, 1989). Instead, the emphasis in reading research shifted to teaching and testing.

Noam Chomsky's (1959) review of Skinner's (1957) *Verbal Behavior* and behaviorism was the harbinger of the third period of research on reading cognitive processes (Rayner & Pollatsek, 1989). This period witnessed the emergence and modification of multiple reading models and theories (e.g., Gough, 1972; Goodman, 1967; Smith, 1971; Kintsch & van Dijk, 1978; Just & Carpenter, 1980; LaBerge and Samuels, 1974; Rumelhart, 1977; Stanovich, 1980; Perfetti, 1985; Anderson & Pearson, 1988; Carrell, 1988; Rayner & Pollatsek, 1989; Carver, 1997; Hoover & Tunmer, 1993).

A major theme of the research on the cognitive processes of reading in this period is characterized by the debate between two views of reading (e.g., Carrell, 1988; Stanovich, 1980; Goodman, 1981). One group of scholars believes that reading is basically a data-driven, bottom-up process (e.g., Gough, 1972; LaBerge and Samuels, 1974). These researchers regard reading primarily as a process of decoding the text from

the smallest unit, i.e., letters, to the construction of the meaning of phrases, clauses, and the whole text. Another group of researchers emphasizes the active role played by the reader, in generating expectations, sampling, and confirming expectations (e.g., Goodman, 1967; Smith, 1971). They also believe that reading activities are primarily influenced by the reader's prior knowledge and experience. Still another group of researchers take an interactive perspective of reading (e.g., Kintsch & van Dijk, 1978; Just & Carpenter, 1980; Stanovich, 1980). They believe that reading is the interaction between reader variables and text variables. Readers make use of their linguistic and background knowledge at different levels of reading processes to construct the meaning of texts. Furthermore, they emphasize "rapid context-free word recognition" (Stanovich, 1980: 32) in developing skillful reading ability.

In addition to the research on the cognitive processes of reading, some researchers during this period focused on the conceptualization of reading ability. For example, Hoover & Tunmer (1993) proposed that reading ability is the product of word recognition abilities and comprehension abilities (measured by listening comprehension tasks). Since only two components are involved in reading ability conceptualization, Hoover & Tunmer labeled it "The Simple View of Reading". Similarly, based on a large number of assessment data, Carver (1997) developed a model of "rauding" efficiency in which rauding is the combination of reading and auding (listening). Rauding efficiency is an estimate of reading ability. In his view, *rauding accuracy* and *rauding rate* are the two components that contribute to rauding efficiency. Rauding accuracy is comprised of

verbal knowledge and listening ability, while rauding rate is comprised of cognitive speed and listening comprehension. Although models focusing on reading ability are closely related to the research questions of the present study, those L1 reading models with a component of L1 listening comprehension ability contribute little to the estimation of L2 reading ability (Urquhart & Weir, 1998; Grabe, 2009) because oral comprehension ability has not typically developed before L2 readers begin to learn how to read.

Despite the lack of explanatory power of some L1 reading models with a listening comprehension component, other L1 reading theories and models, for example Kintsch & van Dijk (1978) and Just & Carpenter (1980), have had a profound influence on L2 reading scholars (e.g., Urquhart & Weir, 1998; Grabe & Stoller, 2002; Koda, 2005). Kintsch & van Dijk (1978) aimed to describe the mental operations in discourse comprehension and the production of recall and summary of the text. Regarding discourse comprehension, Kintsch & van Dijk (1978) used *propositions* as the basic unit of comprehension. A proposition is comprised of a predicate, or relation concept, and one or more arguments. The relations among the propositions, explicit or implicit, with the interplay of context-specific and general knowledge contribute to the development of discourse comprehension. The cyclical construction of propositions is constrained by working memory. Van Dijk & Kintsch (1983) also made a distinction between the construction of a *textbase* and a *situation model of interpretation*. The former refers to the semantic representation of the discourse in memory, while the latter refers to "the

cognitive representation of the events, actions, persons, and in general the situation, a text is about" (p. 41).

Just & Carpenter (1980) presented another L1 reading model that has greatly influenced L2 reading research. Their model consisted of three major parts, i.e., the execution of reading processes, mediating function of working memory, and long-term memory. They conceptualized reading as the coordinated execution of reading processes, which included word encoding, lexical access, assigning semantic roles, and connecting the information in a sentence to previous sentences or former knowledge. The long-term memory is the storehouse of knowledge that provides procedural knowledge to execute the processes. The working memory mediated the long-term memory and reading processes.

Another theme in reading research is the incorporation of the socio-constructivist perspective in reading research since the 1980s. The core research interest of this perspective rests with the construction of meaning. Socio-constructivist scholars (e.g., Bruner, 1981; Heath, 1981, 1983; Gee, 2001; Schallert & Martin, 2003) believe meaning, including textual meaning, is socially, institutionally, and culturally situated. For example, Heath (1981, 1983) conducted ethnographic research on three different communities: Trackton, Roadville, and Maintown. She found that even the meaning of reading, or the construct of reading, had different meanings to members in different communities. In sum, the socio-constructivist perspective of meaning construction has provided a new lens through which to study literacy.

Research into L2 reading has been largely built upon the theories of L1 reading. However, L2 research has also been expanded by its unique features. The following section will discuss research questions that pertain to the L2 and the ways in which L2 reading researchers conceptualize L2 reading and reading ability.

1.3.2 Conceptualizing L2 reading and reading ability

The construct of L2 reading can be illustrated by examining the differences between L1 and L2 reading. Grabe & Stoller (2002) present seven linguistic and processing differences.

First, most L1 readers possess a large oral vocabulary size before they begin to read. They have acquired most of the basic grammatical structures by this time, as well. However, L2 readers generally do not possess these linguistic resources when they begin to read in L2.

Second, adult L2 readers typically mostly have greater metalinguisitic and metacognitive awareness than L1 children. They have learned how to develop reading skills and the strategies that facilitate learning. Third, compared with L1 readers, L2 readers have a lower amount of exposure to L2 print. The fourth difference involves the distance between the L1 and L2. For example, the difference between Spanish and French is smaller than that between English and Chinese.

The following three differences concern the transfer of L1 proficiency to L2 reading. Since most L2 learners have developed L1 proficiency to varying levels, the role played by L1 proficiency poses an issue in L2 reading research. Some believe that skills

transfer has a positive effect, while others believe it might interfere with L2 development. Still others focus on the condition that renders transfer possible. Clarke (1979) and Alderson (2000) believed that in order to make the transfer of L1 to L2 reading possible, L2 readers must cross a linguistic threshold. Although positive transfer is a popular notion, Anderson (1995) contended that transfer of L1 proficiency might be counterproductive. A related concern involves interactions between two languages. The L2 might have a different effect on various reading processes (Segalowitz et al., 1991), but this issue has scarcely been studied.

In conclusion, the differences between the L1 and L2 and the unique issues related to L2 reading research might provide a broader perspective of L2 reading. In addition, L2 reading scholars have also presented analyses of the processes and components of L2 reading with different emphases.

Urquhart & Weir (1998) synthesized and expanded the L1 reading models of Just & Carpenter (1980) and Kintsch & van Dijk (1978), which are introduced in the previous section. Urquhart & Weir incorporated a *Goalsetter* and *Monitor* into their model. The monitor is controlled by the goalsetter, which refers to the overall reading goals. The monitor serves as the mediator between reading processes and reading resources (working memory, long-term memory, and background knowledge). Reading processes consist of extracting the physical features of a text, encoding words and accessing lexicon, parsing syntactic structures, integrating with representations of previous texts, and obtaining subsequent input by moving the eyes. They divided reading goals into five

types: careful global reading, scanning, skimming, search reading, and browsing. The reading processes are adjusted via the monitor that is controlled by the goalsetter.

Lower-level processes	Higher-level processes
• Lexical access • Syntactic parsing • Semantic proposition formation • Working memory	• Text model of comprehension • Situation model of reader interpretation • Background knowledge use and inferencing • Executive control processes

Figure 1.1 Reading processes that are activated when we read

(Grabe &Stoller, 2002, p.20)

Grabe & Stoller (2002) defined reading as "the ability to draw meaning from the printed page and interpret this information appropriately" (p. 9). They conceptualized L2 reading by first illustrating the processes and components involved in reading and then presenting the differences between L1 and L2 reading. As shown in Figure 1.1, they grouped the processes into lower-level processes and higher-level processes. The former refer to the more automatic linguistic processes, including lexical access, syntactic parsing, semantic proposition formation, and working memory activation. In the higher-level processes, readers integrate information from the text and interpret it with the use of background knowledge and inferential competence. They outline four higher-level cognitive processes, namely, the text model of comprehension, the situation model of reader interpretation, background knowledge use and inferencing, and executive control processes.

Koda (2005) elucidated the essential components in L2 reading from a cross-linguistic approach. In addition to the analyses of individual components, she discussed the influence of L1 variations at the levels of word recognition, lexical organization, sentence processing, and text structure. She also emphasized the development of strategic reading.

In conclusion, L2 reading and reading ability have been illustrated from L1 reading, the differences between L1 and L2 reading, and L2 reading scholars' analyses of L2 reading processes and components.

1.4 ASSESSING READING ABILITY

Reading ability has been widely measured despite its inherently complex nature. Reading tests differ in their purposes and intended uses — features that also guide task design and determine the properties of a test.

1.4.1 Purposes and models of reading assessment

Reading assessment generally serves three main purposes: administrative, classificatory, and diagnostic. Administrative decisions such as funding allocations and policy decisions are often required to be based on test scores due to the requirement for accountability in modern education. Some tests such as the TOEFL are used for the purpose of classification or selection, i.e., university admissions, academic program placement, and graduation qualifications. Diagnostic reading assessments aim at identifying the sources of reading difficulties experienced by underachieving individuals.

The outcomes of this kind of assessment can be used to guide the types of reading instruction provided to the target group or individuals.

Administrative- and classificatory-oriented reading measurements usually refers to large-scale standardized tests, while diagnostic reading measurement is typically in the form of classroom-based performance assessment. With respect to the interpretation of testing scores, large-scale standardized tests are categorized into norm-referenced and criterion-referenced assessment. In a norm-referenced test, a representative group of students, or the norm group, is given the test prior to its availability to the public. The scores of the students who take the test after publication are then compared to those of the norm group. Therefore, test takers' scores indicate their relative standing in reference to the norm. By contrast, in a criterion-referenced test, levels of performance, or a set of criteria, are pre-determined and described. Student performance is then compared with the specified criteria, and the scores refer to the level that they attain.

Large-scale standardized tests are generally either norm-referenced or criterion-referenced. However, the CET has been characterized as a norm-referenced criterion-related test (Yang & Weir, 1998; Yang, 2000; Jin & Yang, 2006). In the case of the CET, the norm was composed of about 10,000 undergraduates from six top universities in China, and the criteria are the language competency requirements described in *The National College English Teaching Syllabus*. From a psychometric perspective, the CET committee adopts an atypical approach to interpreting the scores. It is neither "norm-referenced" nor "criterion-referenced" but "norm-referenced criterion-related". This

property of the CET brings challenges to the interpretation of the CET scores, because a test-taker's score should be interpreted in two ways. First, the score is interpreted as the relative standing of the test-taker in the norm, or the percentile of test-takers that are higher or lower than him or her. Second, the score is interpreted as whether the test-taker has reached the required EFL level described in *The National College English Teaching Syllabus*. However, a detailed discussion of the appropriateness of the CET attributions goes beyond the scope of the current study.

1.4.2 Qualities of L2 reading assessment

Bachman & Palmer (1996) contend that the most important quality of a test is its usefulness, since the primary consideration in developing any language test is the intended use of the test. They further pose a formula for test usefulness as follows:

Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality (Bachman & Palmer, p.18).

In their formula, reliability refers to the degree of consistency of a given measurement. Construct validity concerns the meaningfulness and appropriateness of the interpretations that are made on the basis of test scores. Authenticity is defined as the degree of correspondence of the characteristics of a given language test task to the features of a target-language use task. Interactiveness refers to the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task. Impact refers to the influence of a test on society, educational systems, and individuals such as test-takers and

teachers. The last test quality, practicality, concerns the extent to which the demands of the particular test specifications can be met within the limits of existing resources.

Although six components are included in Bachman & Palmer's (1996) formula, these individual test qualities are not supposed to be evaluated independently. Each specific testing situation determines how to prioritize the components. For example, in classroom-based L2 tests, authenticity and interactiveness are usually prioritized over reliability. However, in large-scale standardized tests, reliability and validity are the most fundamental considerations with respect to measurement.

Because construct validity is the core interest of the present research, it necessitates further discussion. Cronbach & Meehl's (1955) landmark analysis of validity is viewed as the traditional validity theory. Cronbach & Meehl categorize validity into four types: predictive validity, concurrent validity, content validity, and construct validity. The first two types, predictive and concurrent validity, together comprise criterion-related validity, because both are interested in some criterion that the test intends to predict. Content validity refers to the extent to which a testing sample represents a universe in which the investigator is interested. Construct validity concerns the degree to which a test measures a theorized psychological construct, or to what extent that the intended construct accounts for variance in test performance. In psychology, a hypothetical construct is an explanatory variable which cannot be directly observed but can be indicated or manifested by groups of functionally related behaviors, attitudes, processes, and experiences.

Messick's (1989) theory posits a broader definition of validity. According to him, "Validity is an integrated evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriate of inferences and actions based on test scores" (Messick, p. 13). Messick's view of validity emphasizes that various sources of evidence, empirical and theoretical, should be gathered to support the interpretations of test scores. High correlations of one test with other validated tests (concurrent validity evidence), high correlations of one test with test-takers future academic performances (predictive validity evidence), and full operationalization of a theorized concept (theory-based validity evidence) are all evidence that can be used to justify the interpretation of test scores.

The fundamental difference between Cronbach & Meehl's (1955) and Messick's (1989) conceptualization of validity is that the former divides validity into separate types while the latter emphasizes a unitary view. Although researchers today generally embrace Messick's validity theory, the traditional notion of validity can still shed light on our understanding of validity. It is reasonable to hold a unitary view of validity while simultaneously deeming the various types of validity as dimensions or facets of a unitary concept.

1.5 A VALIDITY ARGUMENT FOR READING ASSESSMENTS

While validity is used to describe a property of a test, validation refers to a concrete process. Cizek (2008) wedges Messick's (1989) definition of validity into his understanding of validation, which is as "an ongoing endeavor in which various sources

of evidence are gathered, synthesized, and summarized to arrive at 'an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interferences and actions based on test scores'" (p.398).

Given that the validity of a test cannot be proved but can only be evaluated by various sources of evidence, studies on validation generally adopt an interpretative argument approach or validity argument approach (e.g., Kane, 2002; Mislevy et al., 2002, 2003; Chapelle et al. 2008). This validation approach in educational measurement is an application of Toulmin's (1958, 2003) philosophic work on logical reasoning, in particular, on informal or practical arguments. Different from formal arguments, in which premises are taken as given, the assumptions between observation and claim in informal arguments are generally not explicit. A case in point is that the assumptions, or "warrants" in Toulmin's terms, between the observation of a high score in the CET reading section and the claim that the test-taker has a high reading ability are implicit. The goal of a validity argument is to find evidence to back or to refute the assumptions.

Figure 1.2 depicts Toulmin's (2003) layout of arguments. The letter D stands for data, or the facts that are appealed to as foundation for a claim; C stands for a claim or a conclusion whose merits are under examination; W stands for warrants, or practical standards of argument, which are usually implicit in informal arguments.

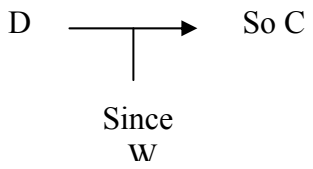


Figure 1.2 The layout of interpretative arguments (Toulmin, 2003, p.92)

To increase the plausibility of an argument, the task is to find evidence to back the assumptions, rather than to strengthen the ground on which the argument is constructed. A number of implicit assumptions, or warrants, might be involved in an argument. To return to the former example of the claim that the test-taker has a high reading ability based on the obtained high score in the CET reading section, a variety of assumptions are implied, including a) the score is reliable; the test-taker will get the same score if he or she attends another version of the test, or if the test paper is scored by another person; b) the administration of the CET is standardized, test-takers having the same amount of test time and in the same kind of physical situation; c) the passages and questions in the CET reading section are a good sample of reading tasks; c) the test-taker is very likely to obtain a high score in a different reading test; d) the good performance in the CET reading section is attributed to the person's actual reading ability, rather than lucky guessing, higher IQ than other test-takers, cheating, or background knowledge.

In order to organize various sources of evidence for the validation of L2 reading assessment, Weir (2005) advances a socio-cognitive framework. As shown in Figure 1.2, evidence may emerge from studies of test-taker characteristics, context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity. The

concept of context validity is an expansion of traditional content validity because it includes considerations of task setting and administration setting. Scoring validity is equivalent to reliability.

Weir's (2005) terms Cronbach & Meehl's (1955) concept of construct validity as theory-based validity, which he intends to differentiate this narrow sense of construct validity from Messick's (1989) expanded concept of validity. As shown in Figure 1.2, theory-based validity involves gathering sources related to the executive processes of the construct of reading ability, strategies to monitor reading process, as well as executive resources such as grammatical and textual knowledge and background knowledge about the reading topics. These sources of evidence should be synthesized and summarized to form a judgment of the degree to which reading scores reflect actual variation in reading ability. Importantly, Weir's view of the L2 reading process is very similar to Grabe & Stoller's (2002) cognitive processing perspective of reading. In fact, Weir recommended the use of Grabe & Stoller's processing perspective of reading in theory-based reading validation studies.

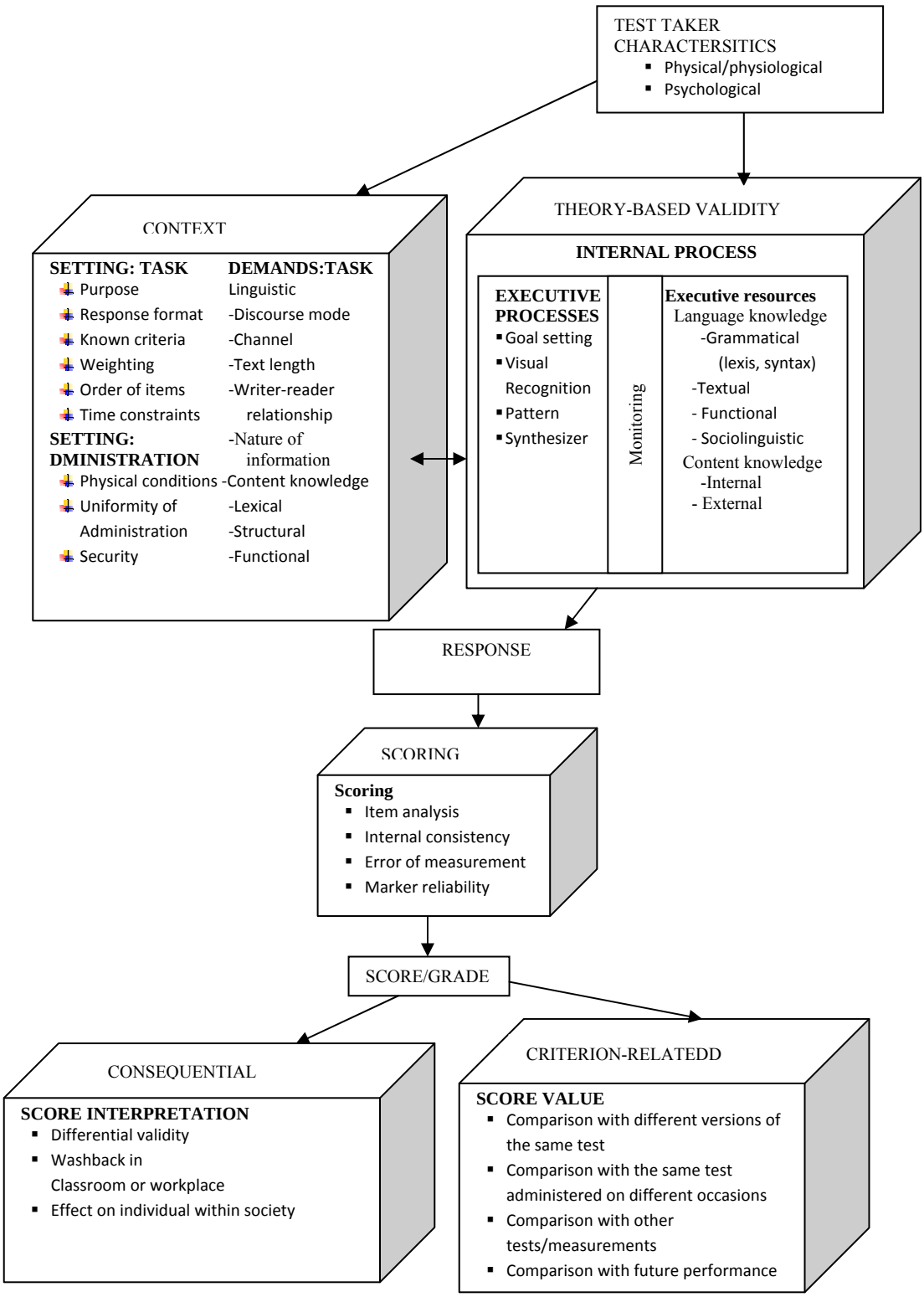


Figure 1.3 Framework for Validating Reading Tests (Weir, 2005, p.44)

1.6 RESEARCH QUESTIONS

Employing an interpretative argument approach (Toulmin, 1958; 2003; Kane, 1992; 2001; Mislevy et al., 2002; 2003; Chapelle et al., 2008), and Grabe & Stoller's (2002) processing perspective of L2 reading ability, as well as Weir's (2005) conceptualization of L2 reading, the present study intends to evaluate the extent to which the CET reading scores can be used to draw inferences about variation of Chinese test-takers' reading ability in English. A finding that reading scores are both grounded in and attributed to the efficiency of reading processing and the high competency of the component skills would provide evidence that test performance on the reading section of the CET reflects the theoretical construct of reading ability. Such a finding would ultimately indicate the degree to which the CET is a valid test of reading ability. In other words, if the underlying abilities of test performance are found to be attributed to theoretical reading processing skills and to monitoring and executive resources (content and language knowledge), rather than to theoretically unrelated processing components (e.g., guessing, topical knowledge, test practice effect), then there would be evidence to suggest that test scores can be used to draw inferences about test-takers' reading ability.

Two interrelated research questions are posed:

1. To what extent are test-takers' performances on the CET reading section attributed to their reading ability?
2. To what extent do test-takers' reading abilities account for their performances in the reading section of the CET?

Figure 1.4 illustrates the structure of the interpretative argument of the present study. The rectangle at the bottom represents the grounds of claims about test-takers' reading ability, which are the scores that are assigned to test-takers. The rectangle on the top indicates claims that are made based on the scores. The claim could be "the test-taker's reading ability is weak" or "the test-taker's reading ability is strong". The claim is valid, or the interpretation of the score is valid, if the warrant that test-takers' performance is attributed to their reading ability is supported. The warrant is represented by the left top rectangle. The research question is to what extent that test-takers' performance is attributed to their reading ability, which is depicted by the rectangle below the warrant. The claim might be refuted if negative evidence is identified, which is not the focus of the present study.

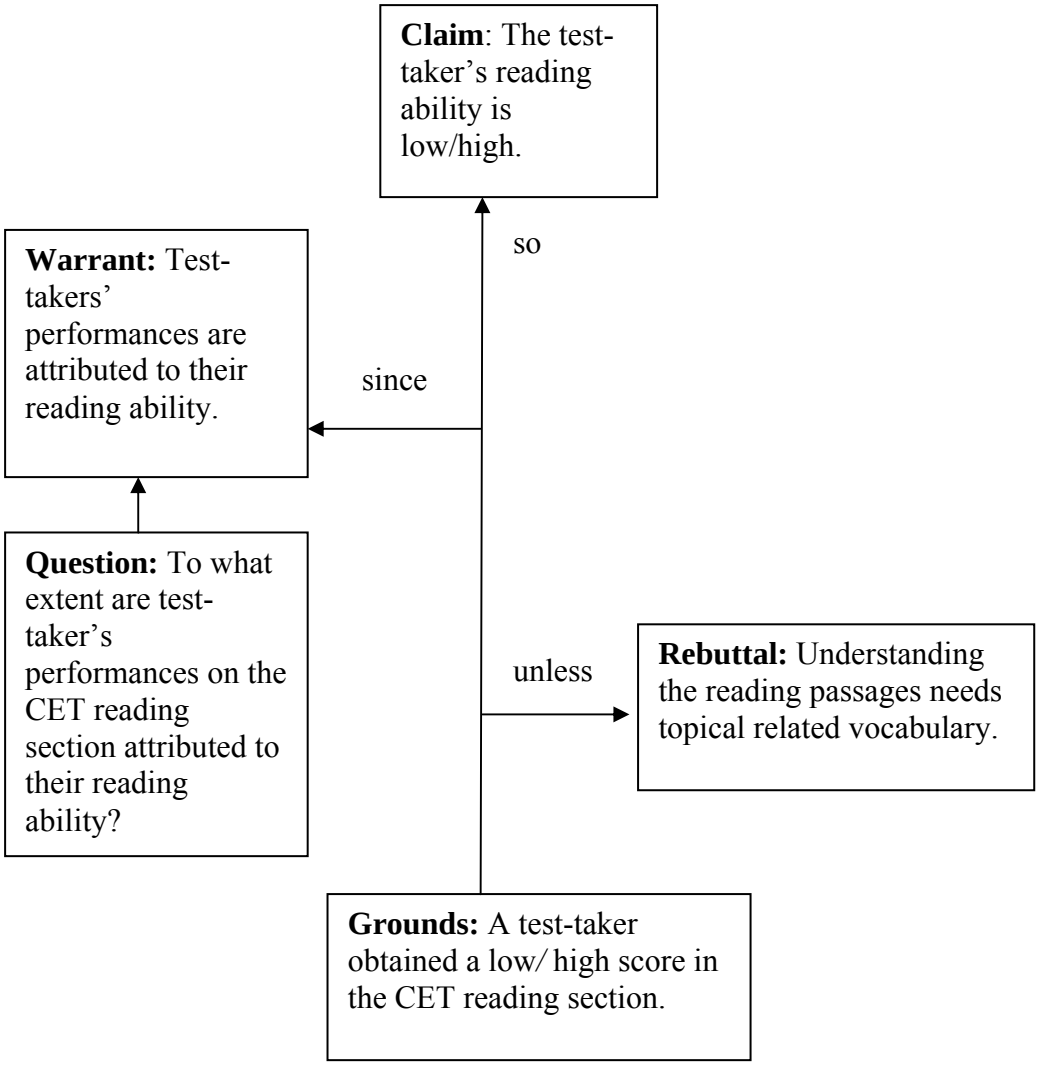


Figure 1.4 The interpretative argument and the research question

(Adapted from Chapelle et al., 2008, p.7)

1.7 COMPONENTS SELECTED TO ESTIMATE READING ABILITY IN THE PRESENT STUDY

Reading scholars generally agree on the essential components involved in reading, although they differ in the analysis of the nature and actual processes of reading,

which as shown in the presentation of section 3 Chapter 1 of this study. The component perspective of reading is also helpful in estimating reading ability (Hoover & Tunmer, 1993; Carver, 1997; Urquarhart & Weir, 1998). The following will introduce the six components used to model reading ability in the present study.

First, there is a general consensus as to the function of word recognition in L2 reading based on the above analysis (LaBerge & Samuels, 1974; Adams, 1994; Stanovich, 1991; Perfetti & Lesgold, 1977, 1979; Perfetti, 1991; Koda, 1996; Koda, 2005; Urquhart & Weir, 1998; Grabe & Stoller, 2002; Weir, 2005). Word recognition is defined as the process of translating a visual display of words into phonological codes and lexical meanings. As shown in Figure 1.1, lexical access is the first process in Grabe & Stoller's (2002) analysis of the processes activated when reading. They use the term interchangeably with word recognition meaning "the calling up of the meaning of a word as it is recognized" (p.20). As illustrated in Figure 1.3, Weir (2005) also includes word recognition in his theory-based construct model of reading, although it is labeled as "visual recognition". Therefore, word recognition was included as a component skill to model reading ability in the present study.

Second, working memory was selected as a component to estimate reading ability. It refers to the ability to actively hold information in the mind when we perform complex tasks such as reasoning, comprehending, and learning. It has been commonly assumed that human minds are limited in terms of the amount of information that can be kept active at any given time (Baddeley & Hitch, 1974; Baddeley, 1986, 2007; LaBerge

& Samuels, 1974; Daneman & Carpenter, 1980; Just & Carpenter, 1992). When we read, various processes such as information in-take, short-term maintenance, and retrieval of information compete for the limited resources of working memory. Thus, the trade-off between processing and storage functions is assumed to be causally related to individual reading ability (Daneman & Carpenter, 1980; Just & Carpenter, 1992; Perfetti et al. 2005). Grabe & Stoller (2002) emphasize the importance of working memory by comparing it to an automobile engine. For the above reasons, working memory will also be chosen as a variable to model reading ability in the present study.

Third, the crucial role of semantic knowledge, or vocabulary, to text understanding has been consistently demonstrated (e.g., Davis, 1968; Carroll, 1971; Anderson & Freebody, 1983; Koda, 1988; Qian, 1999; Pulido & Hambrick, 2008; Pulido, 2009). Verhoven's (2000) study even indicated that vocabulary plays a more important role to L2 readers than to L1 readers.

Fourth, as regards the function of syntactic knowledge in reading comprehension, the results of a number of L2 studies (e.g., Berman, 1984; Barnett, 1986; Alderson, 1993; van Gelderen et al. 2004, 2007; Brisbois, 1995; Taillefer, 1996; Lee & Schallert, 1997; Nassaji & Geva, 1999; Nassaji, 2003; Shiotsu, 2003; Shiotsu & Weir, 2007) have indicated that syntactic knowledge is one of the most important components of L2 reading. Exploring approaches to shortening the International English Language Test System (IELTS), Alderson (1993) reported a correlation of .80 between EFL reading and grammatical ability. He even commented "we have no evidence that other components

are more important (than syntactic knowledge)" (p.218). Enright et al.'s study on the new TOEFL found that the structure section of the old TOEFL correlated highly with the piloted reading section of the new TOEFL (r = .83). Recently, Shiotsu & Weir (2007), comparing the relative importance of vocabulary and syntactic knowledge, found that syntactic knowledge contributes more than vocabulary knowledge to the explanation of variance in reading ability. In sum, semantic knowledge, or vocabulary, and syntactic knowledge will be incorporated in this study as components in a model to predict reading ability.

The fifth component is discourse, or textual, knowledge. This term refers to the knowledge of the features and specific devices that are used to achieve textual coherence. Grabe & Stoller (2002) refer to the application of discourse knowledge to coordinate causal-level meanings and form a representation of the main points of the text as a textual model of comprehension. Other researchers of reading (e.g., Bernhardt, 1991; Grabe, 1991; Alderson, 2000; Koda, 2005) have also analyzed the function of discourse knowledge in reading. Readers' discourse knowledge was therefore used to predict reading ability in the current study.

Metacognitive strategy use in reading was the last component selected. Strategic skills are viewed as an important component of reading ability (e.g., Grabe & Stoller, 2002; Urquhart & Weir, 1998; Koda, 2005; van Dijk & Kintsch, 1983). Strategy use in reading is defined as actions selected deliberately to achieve particular reading goals. Weir (2005) incorporates goal setting and monitoring into the construct of reading. He

explains that reading goals determine the choice of appropriate strategies, and monitoring serves to check the effectiveness of reading performance and strategy use. Grabe & Stoller (2002) designate executive control processes to convey the conception of "the abilities to oversee, or monitor, comprehension, use strategies as needed, reassess and reestablish goals, and repair comprehension problems" (p.28). Some reading researchers (e.g., Brown, 1980; Baker & Brown, 1984; Block, 1986) describe these processes as metacognitive reading strategies. An extensive amount of studies have been conducted on reading to explore the contribution of strategy use to reading comprehension. These studies reveal that the frequency and types of strategy use differ across readers of different reading proficiencies (e.g., Anderson, 1991; Schoonen et al., 1998; van Gelderen et al., 2004; Abbott, 2006; Phakiti, 2008; Plakans, 2009). In the present study, metacognitive reading strategy use was chosen as a component to model reading ability.

Finally, readers' background knowledge was not included in the present study. Although some researchers (e.g., Anderson & Pearson, 1988; Rumelhart, 1980; Carrell, 1988) believe that comprehension involves one's knowledge of the world, its function on reading is not well understood. However, the main reason that background knowledge was not incorporated is that the topics of the CET reading passages were not accessible to the researcher of this study, which rendered the measurement of background knowledge impossible. Even without the component of background knowledge, the confirmatory factor model of reading would not be overly biased. Grabe (2009) divided background knowledge into general knowledge of the world, cultural knowledge, topical knowledge,

and specialist expertise knowledge. Since the participants of the present study came from the same university and from a homogenous culture, there would have been very small variance in their general knowledge of the world and cultural knowledge. Furthermore, the reading passages of the CET do not involve the use of specialist expertise knowledge. There might be some variance in topical knowledge, but the CET committee controls the influence of test-takers' topical knowledge by balancing the four passages among different topics (Yang & Weir, 1998).

In conclusion, six components were selected to model Chinese undergraduates' reading ability, i.e., word recognition, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy use in reading.

1.8 CONFIRMATORY MODELS FOR READING ABILITY

The structure of the six components is the central concern for the following analysis. CFA will be employed because it typically functions to verify a hypothetical factor structure that is based on theories or empirical research. In the present study, the researcher postulates six components based on reading theories and attempts to test whether a relationship exists between the six variables and their underlying reading ability. This type of analysis is a both preliminary and a prerequisite step for the main research questions of this study. If a CFA model can be retained, the relationship between the CET reading scores and reading ability will be examined by a structural equation model. Finally, the degree to which the CET reading score variance is explained by the modeled reading ability will be examined. If CET reading scores reflect variation in

reading ability, then the scores might be used as evidence to support inferences about the test-takers' reading abilities.

Three slightly different models are generated by the present study based on Grabe & Stoller (2002), Weir (2005), and Koda (2005) but none is a strict operation of their analyses. Hypothesized confirmatory model 1 of reading ability (Figure 1.5) is more similar to Grabe & Stoller's conceptualization of reading ability, while Hypothesized confirmatory model 2 of reading ability (Figure 1.6) is an approximate operation of Weir's analysis of executive processes and resources of reading ability. Hypothesized confirmatory model 3 of reading ability (Figure 1.7) agrees more with Koda's (2005) analysis of essential components of reading. It does not categorize the six components into higher and lower processes or into executive processes and resources, as do the models in Figures 1.5 and 1.6, respectively.

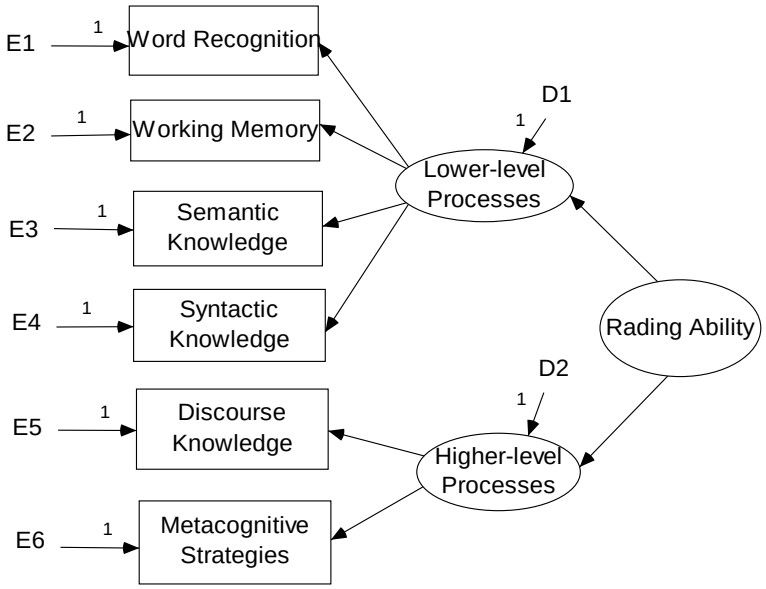


Figure 1.5 Hypothesized confirmatory model 1 of reading ability

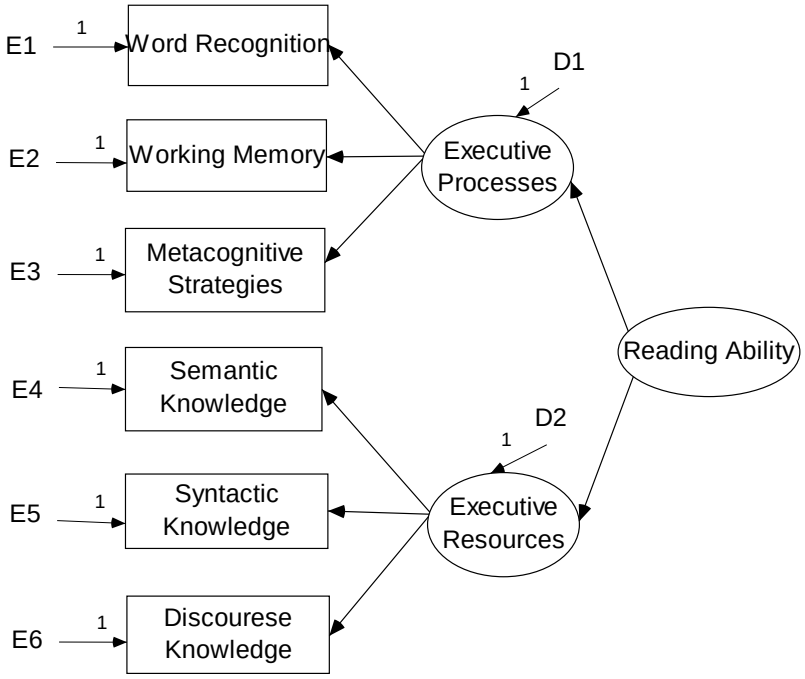


Figure 1.6 Hypothesized confirmatory model 2 of reading ability

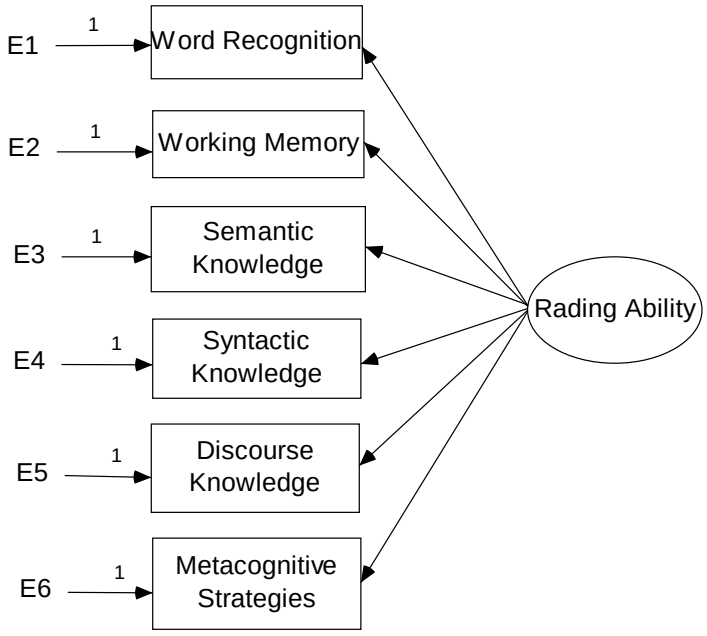


Figure 1.7 Hypothesized confirmatory model 3 of reading ability

In conclusion, this chapter presents the rationale for the present study, followed by the research questions, and the research design. In chapter two, studies on the quality of the CET will be reviewed. Chapter three will introduce the measurement instruments for the six components used to model reading ability as well as the methods for data collection, scoring, and data analysis. Chapter four will report the results of the confirmatory factor analysis of reading ability and its relationship to the test performance on the CET reading section. The final chapter will synthesize and interpret the results of the present study and present a discussion of the implications for understanding the nature of L2 reading and for L2 reading instruction.

Chapter 2 Literature review

This chapter will review 12 validation studies on the CET, which represent all of the studies that examine the quality of the CET to the best knowledge of the researcher. To begin, Yang & Weir (1998), which is the first and most comprehensive examination of the validity of the CET, will be reviewed. Second, studies on the quality of individual sections of this test will be introduced and examined, including three studies on the CET Spoken English Test, one study on listening, one study on reading, two studies on writing, and two studies on translation. Finally, the present literature review will focus on two studies of the consequential validity of the CET.

2.1 YANG & WEIR'S (1998) VALIDATION STUDY OF THE NATIONAL COLLEGE ENGLISH TEST

Yang & Weir's (1998) study was published after the CET had been administered for a decade. The study was a three-year joint project between the CET committee and the Center for Applied Linguistic Studies, University of Reading, UK. Yang & Weir examined the CET construct validity, content validity, concurrent validity, and face validity.

First, regarding the CET construct validity, Yang & Weir (1998) analyzed different theories about the construct of foreign language proficiency and corresponding test methods. Weighing the pros and cons of different test approaches, the CET committee decided to adopt a hybrid method that incorporated structural, integrative, and communicative test approaches.

Second, in regard to the content validity, the study first reported the procedures that the committee adopted for the generation of random samples of test items. The study concluded that the test papers in the previous years had fully covered the content areas as outlined in the test specification. This part of the comprehensive study of Yang & Weir (1998) provided positive evidence in favor of the content validity of the CET.

Third, the study examined the scoring reliability of the writing section of the CET. Based on an analysis of 230 raters' scores of 10 essays the results showed that there existed no significant differences between experienced and novice raters. This portion of the study presented positive evidence for the scoring reliability of the CET writing section.

Fourth, to examine concurrent validity, 660 participants took the CET and the Japanese Society for Testing English Proficiency (STEP) examination. The results indicated that the Pearson correlation between the participants' performance on these two tests was .67.

Fifth, the study compared the rank orders of the CET scores and teachers' evaluations of participants' English proficiency, and the resulting rank order correlation was .70.

Finally, regarding the face validity, the study incorporated a survey of 3149 participants that asked them whether they believed the CET test items were appropriate for measuring college students' English proficiency. The results revealed that overall the participants, including college students, teachers, and companies that required

prospective employees to speak English, evaluated the test items positively. However, some participants suggested that the CET should also integrate a speaking English test and should add translation items.

In summary, Yang & Weir (1998) is a comprehensive validation study that covers various types of validity evidence. Furthermore, it is the only study that has examined the construct validity of the entire CET; other studies only focus on one section of the CET.

However, judging from Weir's (2005) framework of validity evidence, as shown in Figure 1.2, Yang & Weir's 1998 study has neglected evidence related to theory-based evidence, which is the core of the present study.

2.2 STUDIES ON THE CET SPOKEN ENGLISH TEST (CET-SET)

The CET-SET, which was administered for the first time in 1999, adopts a face-to-face format. The test consists of three parts: a warm up (5 minutes), individual presentation and discussion (10 minutes), and further questions (5 minutes). The purpose of the test is to measure test-takers' oral communicative English ability. Two raters assess a group of three candidates simultaneously (and on rare occasions, four candidates).

Three studies on the CET-SET are reviewed in this section: Jin (2000), He & Zhang (2008) and He & Dai (2006). Jin surveyed 358 test-takers' and 28 raters' opinions about the format, length, and validity of the test. Among the 358 student participants, 84.6% considered the format, which utilizes two raters to assess three candidates at a time, "very good" and "good", while 96.3% of the raters held this opinion. With regard to the testing time, 64.5% of the test-takers thought 20 minutes as an appropriate amount of

time, 34.5% considered this amount of time to be too short, and only two thought it was too long. Among the 28-rater participants, 27 believed that 20 minutes for the test was appropriate and only one thought it was too short. As to their opinion about whether the test evaluated the test-takers' oral communicative English ability accurately, 78.7% of the student participants held a positive opinion of the rating system, while all the raters believed that the ratings were accurate appraisals of the test-takers' real English speaking ability. Overall, Jin's study provides positive evidence to the validity of the CET-SET.

He & Zhang (2008) examined the reliability of the CET-SET by investigating and modeling sources of error variance within the framework of the many-faceted Rasch model (MFRM). Utilizing the Facets Version 3.58 software to analyze the raw scores assigned by 18 raters to 529 test-takers oral performances, the study detected statistically significant differences among all facets including rater severity, task difficulty, rating criteria (accuracy and range, discourse length and coherence, and appropriateness), and rating scales. Although the major purpose of the study was to explore the use of MFRM and the Facets software for detecting measurement error sources and for generating fair scores for each task-taker, the results implied that the reliability of the CET-SET should be improved.

He & Dai (2006) focused on the validity of the CET-SET. Communicative language ability, which is the underlying construct of the CET-SET, was operationalized by eight types of functions described in the CET-SET syllabus: 1) (dis)agreeing, 2) asking for opinions and information, 3) challenging, 4) supporting, 5) modifying, 6)

persuading, 7) developing, and 8) negotiating meaning. Based on 48 group discussions and a 170,000-word corpus of test performance, the study found that (dis)agreeing accounted for 49.5% of the total number of interactional occurrences. The other types of communicative functions were not appropriately elicited from test-takers. The results revealed that the testing tasks are not properly designed to measure the multi-traits of test-takers' oral English communicative ability. The results also provided negative evidence for using the CET-SET scores as a measure of students' oral English communicative ability.

2.3 STUDIES ON THE CET LISTENING SECTION

Unlike other posteriori studies of the CET, Pan (2003) focused on providing a priori validity evidence for the listening section of the CET. Based on Bachman & Palmer's (1996) principle of test construction and the analysis of the task format, as well as the characteristics of some influential large-scale tests (TOEFL, IELTS, PET, TEEP), she focused on the construction of a descriptive system of listening test tasks to provide guidelines for the design of CET listening tasks and shed light on the study of the content validity of the listening section.

2.4 STUDIES ON THE CET READING SECTION

Jin & Wu (1998) utilized an introspective approach, or think-aloud protocol, to explore whether test-takers' employment of reading skills, which were judged by their reading behaviors, matched test designers' expected reading operations on the CET

reading section. Fifty-one students at various language proficiency levels participated in the study. Recording materials of 40 participants were analyzed. Expected reading operations (ERO) were categorized according to nine types:

ERO 1: reading for literal meaning

ERO 2: reading for implied meaning

ERO 3: reading for the main idea

ERO 4: reading to find the author's attitude and opinion

ERO 5: understanding the contextual meaning of a word

ERO 6: understanding the meaning of a sentence

ERO 7: understanding the discourse structure

ERO 8: skimming for the main idea

ERO 9: scanning to find specific information

Based on the above nine reading operations, Jin & Wu (1998) outlined nine expected corresponding reading performance (ERP) types for readers.

ERP 1: Students are expected to locate the target sentences, read them, and answer questions pertaining to ERO 1.

ERP 2: For ERO 2 questions, students are expected to read related sentences or paragraphs carefully and draw inferences. Students are expected to return to the text to double check their answers.

ERP 3: For ERO 3 question types, students are expected to focus on locating topic sentences, usually at the beginning or end of paragraphs.

ERP 4: For ERO 4 reading questions, readers are expected to read the whole text, focus on some key sentences, and draw inferences.

ERP 5: For ERO 5 questions, students are expected to locate the target sentences and read sentences that come before and after them.

ERP 6: For ERO 6 questions, students are expected to understand the content words and the function words of the target sentences.

ERP 7: For ERO 7 questions, students are expected to understand sentences as well as text structure.

ERP 8: For ERO 8 question types, participants are expected to read fast and use skills of ERO 3 and ERO 4 skills.

ERP 9: For this last type of question types, students are expected to locate target words or sentences quickly and use the ERO 1 skill.

Jin & Wu (1998) also listed five strategies that they do not believe to be components of the conceptualization of the construct of reading ability. These five strategies are 1) to eliminate distracters by reasoning and not by reading comprehension, 2) to use background knowledge and common sense and focus on the question stems to answer reading questions, 3) to guess, 4) to locate the answers to questions by the sequence of the questions, such as to find the answer to question 1 in the first paragraph and the answers to the last question in the last paragraph, and 5) to focus on finding matches between phrases in the question stems or choices with phrases in the text.

The results indicated that among the 289 cases of CET-4 correct answers, the participants employed expected reading skills 258 times (89.3%), while the participants did not employ expected reading skills, only 31 times (10.7%). Among the 111 cases of incorrect answers, participants utilized the expected reading skills 20 times (18.0%), and participants did not use expected reading skills, 91 times (82.0%). The authors concluded that the CET reading tasks could effectively measure test-takers' reading ability because correct answers were largely based on the match between test-takers' employment of reading skills and test designers' expected reading operations.

Jin & Wu's 1998 study intended to explore theory-based validity evidence for the interpretation of the scores on the reading section of the CET, which is similar to the goal of the present study. However, Jin & Wu's study was conducted more than a decade ago. The CET reading section underwent tremendous changes in 2005. As reported in Chapter one, the original four careful reading passages with 20 total questions were decreased to two careful reading passages with 10 total questions. A fast reading (10%) and a new form of reading task — a cloze with a word bank — were introduced. However, no study on the reading section has been carried out since then.

2.5 STUDIES ON THE CET WRITING SECTION

Wang (2004) compared the CET essay-rating reliabilities using the newly implemented online marking system and the traditional conference-marking method. Fourteen CET raters scored a random sample of 1,341 essays, 20 of which were marked by all 14 raters. The results showed that the standard deviation of the scores was higher

using the online marking system than by adopting the traditional conference method (2.34 versus. 2.22, respectively), and the inter-rater reliability was higher using the new marking system (.84 versus .77). This result provides positive evidence in favor of the new scoring mode.

Using the multi-faceted Rasch measurement computer program, Wang et al. (2006) further examined the severity and consistency of the 14 raters using the online marking system and the conference method reported in Wang (2004). The results revealed that the scores had a wider spread with the online system (9 logit units) than with the traditional method (5 logit units), which was consistent with the results reported in Wang (2004). The average severity of the 14 raters was not distinguishable in the online system, while the severity of the raters was not consistent in the conference setting. Therefore, the on-line scoring method distinguishes test-takers' writing better, and the raters are more consistent in severity when they score essays on-line. This follow-up study of Wang (2004) also provides positive evidence for the new scoring method.

2.6 STUDIES ON THE CET TRANSLATION SECTION

Huang & Liu (1996) compared the CET task types and the College English Course Syllabus and proposed that translation should be incorporated into the CET to improve the content validity of the test. That same year, translation was built into the CET (a perhaps coincidental and overly expeditious response to their concern). The task required test-takers to translate sentences that were underlined in the reading passages into Chinese. In order to explore test-takers' responses to the new task type, Huang et al.

(1996) conducted a survey of 58 participants. They found that 50% of the participants did not think selecting sentences from the reading passages was appropriate, 58.6% responded that finding the target sentences was time consuming and inconvenient, and 63.8% did not have enough time to complete the task. Huang et al. suggested that translation tasks should not overlap with reading because testing items should be independent to improve content validity.

2.7 STUDIES ON THE CONSEQUENTIAL VALIDITY OF THE CET

Han et al. (2004) examined the washback of the CET by surveying 1,194 teachers from 40 universities in China. Negative comments on the CET from the participants characterized the findings of Han et al.'s study. They found that 62.1% of the participants did not believe that the CET could promote English teaching or help students master English linguistic knowledge. The participants believed that the CET interfered with normal classroom teaching to varying degrees: 41.4% for severe interference, 48.3% for some interference, and 10.3% for no interference. Interestingly, despite the negative impact, 70% of participants believed that the CET should not be abolished from college English education.

Gu (2005) examined the CET washback on the EFL teaching and learning of Chinese undergraduates. Various research methods were employed, including classroom observation, questionnaires, interviews, tests, and CET scores. With a total of 4,500 participants, Gu found that the positive impact of the CET outweighed the negative effects. Some positive washback effects included 1) the CET promoted the

implementation of the National College English Course Syllabus, 2) administrators attached greater importance to College English teaching, and 3) the CET motivated teachers and students in their teaching and learning. One negative impact of the CET was the fact that teaching materials could not be finished especially in the fourth semester, due to the time needed to prepare for this test.

Table 2.1 Validation studies on the CET

Study	**Type of validity evidence**	**Positive/ Negative**
Yang & Weir (1998)	Construct, content, scoring reliability concurrent, & face validity	Mainly positive
He & Zhang (2008)	Scoring reliability	Negative
He & Dai (2006)	Theory-based construct validity of the Spoken English Test	Negative
Pan (2007)	Item characteristics in the listening section as evidence for content validity	Priori study
Jin & Wu (1998)	Construct validity	Positive
Wang (2004)	Writing scoring reliability of a new method	Positive
Wang et al. (2006)	Writing scoring reliability of a new method	Positive
Huang & Liu (1996)	Content validity of the CET	Negative
Huang et al. (1996)	Content validity of the CET	Negative

Jin (2000)	Consequential validity of the Spoken English Test	Positive
Gu (2005)	Consequential validity of the CET	Positive
Han et al. (2004)	Consequential validity of the CET	Negative

A number of themes have emerged from the above literature review. First, there is a severe dearth of studies on the validity of the CET. Only 12 studies have been identified, which is surprisingly meager given the large-scale and high-stakes nature of the CET, and the 12 studies examined different aspects of the CET. Millions of Chinese undergraduates have participated in the test as stated in Chapter one, and thousands of EFL teachers and college administers' lives have been influenced by the CET. A limited number of studies are unable to provide adequate evidence regarding the degree of justification for drawing inferences based on the CET scores. Furthermore, the CET has been administered for about a quarter of a century with two sittings per year, and the number of administrations of the CET stands in sharp contrast with the small number of studies on the validity of the CET.

Second, owing to the small total number of studies on the CET, some types of validity evidence have been neglected or even ignored. As revealed in Table 2.1, there is only one study of the criterion-related validity of the CET: in Yang & Weir's 1998 comprehensive validity study. But no one has explored the content validity or theory-based validity of writing. Nor are there any studies that have examined the theory-based

validity of the listening section of the CET. Finally, no study has examined the content validity of the listening section, although a new test format, compound dictation, has been added since 1999. As regards the reading section, no research has been conducted on the content validity, even though the reading tasks on this test have changed dramatically since 2005. The new version has incorporated two new forms of reading tasks, i.e., fast reading and cloze with banked words, while the old version had only careful reading questions. Overall, evidence is insufficient for each type of validity evidence, each component skill of language proficiency, and for the language proficiency of the CET as a whole.

Finally, the 12 reviewed studies do not help different stakeholders to form an informed judgment of the CET. As revealed in Table 2.1, the findings of seven of these studies feature positive evidence, while the results of the remaining five studies are characterized mainly by negative evidence.

In conclusion, the above themes that have been found in the literature review indicate that further evidence for the validity of the CET needs to be explored. Importantly, the sole study that focused on the reading section of the CET, i.e., Jin & Wu's 1998 was conducted more than a decade ago. Considering the significance of reading to foreign language proficiency, and the paucity of studies on the reading section of the CET, as well as the dramatic change in the reading section since 2005, further studies on the CET reading section are in urgent need. The present study will focus on the exploration of the theory-based validity of the CET reading section. In particular, this

study will examine whether the CET test-takers' performances on the CET reading section are underlined by their reading ability, which is indicated by reading processes, linguistic knowledge, and metacognitive strategy.

Chapter 3 Research methods

The present study intends to evaluate the extent to which reading ability accounts for test-takers' performances on the CET reading section so as to examine the construct validity of the reading tasks on the CET. The present study embarks on modeling reading ability by utilizing six theoretically proposed and empirically evidenced components, i.e., word recognition, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive reading strategies. A confirmatory model for reading ability with the best model-fit indices was selected. Finally, a structural mode with the scores in the CET reading section was analyzed, and the linkage between reading ability and test performance on the CET reading section was examined.

This chapter will first introduce the method for participant recruitment and the desgin of the six instruments that are used to meaure the components for modeling reading ability. Second, the procedures for data collection will be laid out. Finally, methods of data analysis will be illustrated, including item scoring, variable index generating, and statistical procedures.

3.1 Recruitment of participants

A participant recruitment advertisement (Appendix A) was posted around the campus and on the website of a large comprehensive university in central China one week after the national CET-4 examination was administered. The recruitment flyer provided the information about the purpose of the study, requirements for participation, time, location, and compensation for participation. It also listed three phone numbers, an

email address, and the coordinator's name for prospective participants to make appointment or to obtain further information.

3.2 INSTRUMENTS

Six instruments were employed in the current study to model reading ability: a pseudowords identification task programmed by DMDX, a revised version of Daneman & Carpenter's (1980) sentence reading span working memory test, Meara & Milton's (2002) Yes/No vocabulary tests, the test of syntactic knowledge used in Shiotsu & Weir's (2007) study, Abeywickrama's (2007) discourse knowledge test, and a revised version of Phakiti's (2008) strategy use questionnaire.

3.2.1 The instrument for measuring word recognition

Word recognition refers to the process of translating a visual display of words into phonological codes and lexical meanings. Due to differing research goals, various tasks have been used in the literature to tap sub-processes involved in word recognition: visual processing, orthographic processing, phonologic processing, and semantic processing. The same-different approach is often employed to measure visual processing ability (e.g., Brown & Haynes, 1985; Haynes & Carr, 1990). This approach usually involves asking readers to identify whether a pair of letter strings are spelled the same or differently. For example, readers are shown "will" and "well", or "hgkgj" and "hgkjg", and they are required to tell whether the two letter strings are the same. Other researchers use case distortion to measure readers' visual processing skills (e.g., Akamatsu, 2003).

Orthography processing refers to the use of orthographic information when processing written code to evaluate whether letter strings conform to English orthographic regularity (e.g., Nassaji & Geva, 1999; Nassaji, 2003). Non-word tasks are often employed to isolate orthographic processing skills from general lexical knowledge to partial out, or to remove, sight word effects (e.g., Siegel et al. 1995; Nassaji & Geva, 1999).

Phonologic processing refers to a systematic and rapid translation of letter strings into pronunciations. Although it seems that sounding out words is not involved in silent reading, phonological processing skills have been consistently documented to be causally related to reading proficiency (e.g., Stanovich, 1986; Share & Stanovich, 1995; Perfetti & Zhang, 1995; Torgesen & Burgess, 1998; Koda, 2005). This skill facilitates reading by enhancing information storage in working memory (Kleiman, 1975; Levy, 1975) and affords quick access to oral vocabulary in lexical memory. A variety of methods have been employed to tap phonologic processing abilities, such as asking participants to read pseudowords aloud (e.g., Brown & Haynes, 1985; Haynes & Carr, 1990), having participants tell whether pairs of pseudowords (e.g. *thake*/*thack*) sound the same (Nassaji & Geva, 1999), and having participants identify which of a pair of pseudowords (e.g. *kake/ dake*) sounds like a real word (Stanovich & West, 1989; Bell & Perfetti, 1994).

Some studies use one instrument to assess word recognition. Van Gelderen et al. (2004) and van Gelderen et al. (2007) employed a lexical decision task to tap Dutch native speaking EFL learners' word recognition efficiency. The task consisted of 60 letter

strings. Half of them were authentic common English words, and the other half were orthographically and phonologically possible pseudowords. The participants were asked to decide as quickly as possible whether a letter string was an existing word.

Efficiency based on speed and accuracy has been shown to be a more accurate predictor of reading ability than simple accuracy measures (Carr et al. 1990; Nassaji & Geva, 1999). Two common methods that have been used to calculate the efficiency of word recognition processing involve reaction time and the number of correct responses per minute. An example of the former is Stanovich & West's 1989 study, in which a composite index was computed by averaging the *z*-scores of median reaction time for correct responses and the number of errors on the task. Haynes & Carr (1990) defined efficiency as the number of correct answers per minute.

In the present study, the identification of pseudowords task was utilized to assess word recognition skill. Because the participants have received English instruction for eight years on average, the visual processing task might not be able to differentiate them, so this type task was not included. The decision not to include this type of task is supported by the finding of Nassaji & Geva's 1999 study, in which a letter-naming task comprising 50 items was used to tap visual processing skills and resulted in a very small standard deviation. The maximum correct response reported in that study was 50, while the minimum was 49. Thus the instrument did not elicit sufficient variation in responses.

A pseudowords identification task, adapted from the work of Olson et al. (1985), was employed to assess word recognition processing skills (Appendix E). This task

assesses the word recognition skill because it taps orthographic, phonological, and semantic processing skills. For example, in order to respond correctly to *thair/theer*, the participants would have to know how to parse the two letter strings, and they would also have to know how to translate *th,air,eer* combinations into their corresponding sounds. Finally, they would have to be able to retrieve the meaning of the sound / ðɛr/ from their vocabulary memory.

The entire task consisted of 25 pairs of pseudowords (e.g. *kake/dake, filst/ferst, thair/theer*) (Appendix E). The participants were required to indicate which letter string sounded like a real English word.

The test was administered individually on a computer. Stimuli were programmed by using DMDX version 4.0.4.4 (Forster & Forster, 2003), which is a Windows-based program designed to measure reaction times with millisecond accuracy of the presentation of text, audio, graphic, and video material. The researcher carried out the following three steps prior to data collection.

Step 1: The DMDX package was downloaded from the web site at http://ww.u.arizona.edu/~kforster/dmdx/dmdx.htm.

Step 2: An input script was written in Word and saved in rich text format (.rtf). The item file is composed of a parameter line, instructions for the participants, two practice trials, and 25 test trials (Appendix E).

Step 3: A specification script telling Analyze 5.1.0 how the raw data are analyzed was written in Notepad and saved in an .spc file.

These three steps comprise the preparation conducted prior to data collection. During data collection, the researcher and her assistants explained the instructions orally to individual participants. The participants were told that this task required them to indicate which letter string of a pair (e.g., *kake/dake*) sounded like a real English word. Then, they were asked to do two practice trials on the computer. Two plus signs appeared first on the screen as a warning signal and then disappeared immediately. Next, two letter strings in lower case appeared on either side of the point where the plus signs were. Participants were required to respond by pressing the two shift keys. They pressed the left shift key if they chose the letter string on the left side and pressed the right shift key if they chose the letter string on the right side. At the end of the practice trial, they were asked, "Do you have any questions?" If anyone responded "yes", further explanations were provided; if there were no questions, the participants were asked to press the keyboard of the spacebar to begin the real test.

3.2.2 The instrument for measuring working memory

Working memory is the ability to actively hold information in the mind when we perform complex tasks such as reasoning, comprehending and learning. Various measures have been adopted to tap readers' working memory capacity, including tests of word span, digit span, oral reading span, silent reading span, and listening span. Other methods, such as rhyming, visual matrices, story recall, picture sequencing, and spatial organization, have been employed to fulfill specific research purposes (e.g., Swanson,

1992, 2003). In the following paragraphs, the procedures of the five most commonly used methods are explained.

Word span tests: Common words are grouped into sets of two to seven words. Each level usually has three sets of words. The words are presented orally or visually at the rate of one word per second. Participants are required to recall the words of a set in the order of presentation. The test begins with sets of two words and continues to sets of three words if a participant answers correctly in all three sets of two-word tests. The test ends at the point in which the participant fails all three sets.

Digit span tests: In this type of test, participants are required to recall a series of orally presented numbers that increase in set size.

Oral reading span tests: Participants are required to read a series of sentences aloud at their own pace and recall the last word of each sentence. Unrelated sentences are grouped into two to six sentences. The test begins with sets of two sentences and advances to sets of three sentences if a participant answers correctly in all three sets of two-sentence trials. The test ends at the point in which the participant fails all three sets.

Silent reading span tests: Participants are required to read a set of sentences silently and make a true or false judgment about the statement. The following are examples of these sentences: (1) *Tables normally have four legs.* (2) *The earth moves around the sun*. At the end of a set, the participants have to recall the last word of each sentence. The true-false component is employed to ensure that participants process the entire sentence rather than concentrating only on the final words. Compared with oral reading span tests, silent

reading span tests are more efficient because they can be administered in groups rather than individually.

Listening span tests: These tests are similar to silent reading span tests except that the participants are required to listen to the sentences rather than read them silently.

In the present study, a revised version of Daneman & Carpenter's (1980) silent sentence reading span test was used (Appendix F). This method has been assumed to consume both processing and storage functions of working memory and has been widely used or adapted in L1 reading studies (e.g., Levy & Hinchley, 1990; Geva & Ryan,1993; Bell & Perfetti, 1994; Gottardo, Stanovich & Siegel, 1996; Waters & Caplan, 1996). However, some researchers believe that reading span tests are less appropriate for L2 studies because the measurement involves language-based processing and memorizing and might be confounded with language comprehension ability and the use of digit span tasks (e.g., Carr et al., 1990; Nassaji & Geva, 1999). To overcome this possible weakness of the silent sentence reading span test, simple sentences with common vocabulary were used. Given that the CET test takers' average vocabulary capacity is around 4,500 words, all of the words in the sentences for the working memory task were controlled within two thousand frequency level vocabulary.

In contrast with Daneman & Carpenter's (1980) version which asked participants to recall the last words of the sentences which they had read, this study employed sentences followed by an unrelated word — a method to control the effect of semantic priming (e.g., Conway et al., 2002). The participants were required to write down the

unrelated word in the order of original presentation. Twenty eight sentences, with a length ranging from five to nine words, were grouped into eight sets, which were divided into four levels of two-, three, four, and five-sentence levels. Each level had two sets of sentences. Participants were required to read each sentence silently and make a true or false judgment about the statement. After they finished reading and judging all sentences in each set, the participants were required to write down the unrelated word attached to the sentence in the order of presentation on the worksheet.

Similar to the measurement of word recognition, the working memory test was administered individually through the DMDX software. The input script was written in Word and saved in the rich text file format (.rtf). A data treatment file was written in Notepad and saved as in the .spc format.

During data collection, the researcher and her assistants first explained the procedure to each participant orally and then administered a two-sentence-level practice test before the real task.

3.2.3 The instrument for measuring semantic knowledge

Semantic knowledge, or vocabulary, is indispensible to reading. Similar to word recognition and working memory, semantic knowledge is also a multi-faceted construct that researchers have conceptualized from various perspectives. Richards (1976) assumed that knowing a word means knowing the following: the degree of probability of encountering that word in speech or print, the limitations imposed on the use of the word according to variations of function and situation, the syntactic behavior associated with

the word, the deviations of the word, the network of associations between that word and other words, the semantic value of a word, and the different meanings associated with a word.

Anderson & Freebody (1981) proposed a two-dimensional perspective on vocabulary knowledge: breadth and depth. The former refers to vocabulary size, or the number of words for which the reader knows the basic meaning. Vocabulary depth is defined as a reader's level of knowledge of various aspects of a given word, such as register, the frequency of the word in the language, syntactic properties, morphological properties, pronunciation, and spelling. The meaning of words includes not only the denotative meaning in context, but also the knowledge of connotations, their antonyms, and synonyms.

Chapelle (1998) elucidated three perspectives of the construct of vocabulary and their implications for vocabulary measurement, namely, a trait perspective, a behaviorist perspective, and an interactionalist perspective. Theorists from a trait perspective would define vocabulary in terms of knowledge and relevant processes, which include the following: vocabulary size, knowledge of word characteristics (e.g., phonemic, graphemic, morphemic, syntactic, semantic, and collocational), lexical organization (the way morphemes and words are represented in the mental lexicon) as well as a set of fundamental processes related to lexical access, such as parsing words into their morphonological components and translating them into pronunciations.

A behaviorist perspective of vocabulary knowledge views a score obtained from a vocabulary test as a sample of the test taker's responses to similar stimuli, which can be used to predict the test taker's performance on similar contextual situations. Therefore, the features of context relevant to vocabulary use are an essential component of vocabulary tests. An interactionalist perspective of vocabulary knowledge entails a description of both vocabulary traits (e.g. vocabulary breadth and depth) and context. A test taker's performance on a vocabulary test is viewed as a sign of his vocabulary knowledge traits, and the performance is influenced by the context in which the test task occurs.

Henriksen (1999) proposed a three-dimensional framework of the construct of vocabulary competence, namely, the partial-precise knowledge dimension, the depth of knowledge dimension, and the receptive-productive dimension. Similarly, Milton (2009) believes that a useful way to describe word knowledge is to divide it into receptive or passive knowledge and productive or active knowledge.

Nation (1990, 2001) advanced a very detailed taxonomy of vocabulary knowledge (see 2001, p. 27). Nation first categories vocabulary into three areas: knowledge of form, meaning, and use. Knowledge of form is further subdivided into spoken, written, and word parts. Knowledge of meaning is then further subdivided into forms and meaning, concepts and referents, as well as associations. Knowledge of use is further dividend into grammatical functions, collocations, and constraints on use.

These different perspectives of vocabulary knowledge have underpinned a variety of vocabulary assessment instruments. Some focus on measuring vocabulary size, others on vocabulary depth, still others on productive ability of vocabulary or on using vocabulary in context. In the category of measuring vocabulary size and receptive knowledge of vocabulary, the Yes/No format test developed by Meara and associates (Meara & Jones, 1988; Meara, 1992; Meara & Milton, 2002) has been widely used and studied (e.g., Meara & Buxton, 1987; Huibregtse et al. 2002; Mochida & Harrington, 2006).

The Yes/No test presents readers with a set of words and instructs them to mark the individual words if they know the meaning of those words. To make appropriate adjustments for random guessing, the test items include pseudowords, which are consistent with phonological constraints but bear no meaning. Participants are informed that the test contains nonsense, but not how many nor their location in the test.

Another widely used test is Nation's (1983, 1990) Vocabulary Levels test. Different from the simple format of Yes/No tests, this type of vocabulary breadth test measures students' vocabulary size through word definitions. The whole test is composed of five vocabulary size levels, namely, the 2,000, 3,000 5,000 and 10,000 word levels. At each vocabulary size level are six test items, each comprising six words and three definitions. Students are required to select three of the six words to match the definitions. For example:

a. royal
b. slow

c. original
d. sorry ____ first
e. total ____ not public
f. private ____ all added together

Schmitt, Schmitt, & Clapham (2001) validated and revised Nation's (1983) test, and extended the number of sets in each level from six to ten.

In the category of vocabulary depth measurement, Read's (1993, 2000) Word Associates test is widely used (Qian, 1999, 2002; Qian & Schedl, 2004). A recent version of the test is composed of 40 items designed to measure meaning and collocation of words. Each item has an adjective stimulus, and two boxes which have four words in each. Among the four words in the left box, one to three words can be synonyms to the stimulus word, while among the four words in the right box one to three words can collocate with the stimulus word. Each item has four correct choices from both left and right boxes. The four correct choices can be one from the left box and three from the right box, two from each box, or three from the left box and one from the right box. The following is an example.

Bright

clever	famous	happy	shining

color	hand	poem	taste

Another vocabulary depth measurement, the Vocabulary Knowledge Scale, was developed by Paribakht & Wesche (1993, 1997). The test estimates students' vocabulary knowledge depth on a five-point scale based on their self-rated familiarity with words, ranging from "The word is familiar and I can use it in a sentence" to "The word is not familiar at all." Different from Read's (1993, 2000) instrument which emphasizes words'

definitional meanings and collocations, Parbakht & Wesche (1993, 1997) treats vocabulary knowledge as a dynamic, continuously changing construct.

In the receptive-productive dimension of vocabulary knowledge, Laufer & Nation (1999) devised gapped word completions tasks. For example, test takers are asked to read the sentence with a gapped word of which the first two letters *p* and *h* are provided:

After finishing his degree, he entered a new ph[ ] *in his career.*

Test takers are required to complete the spelling of the word.

Some vocabulary instruments attempt to tap test takers' knowledge of contextualized meaning of vocabulary. Weir et al. (2000) is an example. The test requires students to choose words from a word bank and fill blanks in two different passages that are each approximately 500 words long. This method of vocabulary assessment has been adopted by the CET since 2005.

In the present study, three vocabulary frequency levels, i.e., 4K, 5K, and 6K, of Meara & Milton's (2002) Yes/No test was employed (Appendix G). First, multiple studies have revealed that it has high validity (Meara & Buxton, 1987; Meara, 1996; Mochida & Harrington, 2006). Meara and colleagues reported a moderately strong correlation (around r = 0.7) between performance on the test and other commonly used European ESL multiple-choice tests of vocabulary. Similarly, Mochita & Harrington (2006) found strong correlation between the performance on the test and on the Vocabulary Levels Test (Nation, 1990). Second, given the multiple components of the measurements involved in the study, efficiency entails a major consideration. Compared

with other instruments, the Yes/No test is the most efficient. Each frequency level normally takes shorter than three minutes. Considering the participants' vocabulary level is around 4,500, three frequency levels have been chosen for the present study, i.e., 4000, 5000, and 6000.

3.2.4 The instrument for syntactic knowledge

After lexical information is obtained and stored in working memory, it must be incrementally integrated into larger linguistic units: phrases and sentences. Comprehending sentences entails not just lexical knowledge but also syntactic ability, which is defined as how a reader knows about the way in which words and phrases are combined to form sentences in a language. Without this knowledge, incorrect multiple semantic interpretations cannot be ruled out, attachment of words and phrases might not be appropriated assigned, relationship between nouns and verbs might not be correctly established, relationship between main clause and subordinate clause might not be accurately understood.

The present study employed Shiotsu's (2003) final 32-item version of syntactic measurement (Appendix H). First, the participants' background of the present study is similar to those of the main study in Shiotsu. Participants are both EFL undergraduates. Second, the original 35-item measurement, which has incorporated 20 items from the Test of English for Educational Purposes (TEEP) by Weir (1983: 371-373) and 15 items from TOEFL (Duran et al., 1985), had undergone a content validation study involving 11

L1-English ELT experts and excluded three items resulting in the final 32-item version. A permission letter from Shiotsu has been obtained.

3.2.5 The instrument for discourse knowledge

Discourse knowledge refers to the knowledge of the features and specific devices that are used to achieve text coherence. In the reading literature other terms, such as "rhetorical knowledge", "text structure knowledge", "knowledge of cohesion and coherence", and "formal knowledge", are used to refer to a similar concept. In the present study "discourse knowledge" is used consistently to avoid a possible confusion of terminology.

Researchers have studied and classified the devices used to achieve text coherence of structural organization from various perspectives. Halliday & Hassan (1976), an influential work in studying cohesion in English, identified five major devices to achieve cohesion, namely, *reference, substitution, ellipsis, conjunction,* and *lexical cohesion*. In each major category, different subtypes were also elucidated. For instance, the category of *conjunction* was further classified into devices of *temporal, additive, causal*, and *adversative*.

Focusing on expository texts, Meyer (1975, 1985) identified three primary levels for prose analysis. The first was *micropropositional* level, which is concerned with the way ideas are organized within sentences. The second was *macropropositional* level, which concerns the issue of logical organization and argumentation. The third is the *top-*

level structure or overall organization of the text, which is further categorized into *problem/solution, comparison, causation*, and *description.*

Graesser et al. (2003) discussed a similar set of discourse types. To exemplify the patterns of global organization of discourse, they listed *setting-plot-moral, problem-solution, compare-contrast, claim-evidence, question-answer,* and *argue-counterargue*.

A series of studies on L1 reading have shown that instruction about discourse structure has yielded positive effects on reading comprehension and remembering information from text (e.g. Bartlett, 1978; Geva, 1983; Taylor & Beach, 1984; Meyer et al., 1989; Meyer & Poon, 2001). Meyer & Poon (2001) comparing the effects of nine-hour structure strategy training, interest strategy training, and no training, found that only structure strategy training group showed increased total recall as well as recall of the more important information. The structure training involved helping students recognize various structural patterns in texts (e.g., comparison and contrast, problem/solution, causation).

Although scarce in number, studies on L2 reading have also demonstrated the positive effect of text structure instruction on reading retention and comprehension. Carrell (1985) conducted an experimental study on 25 intermediate-level ESL students, with 14 in the experimental group and 11 in the control group. The 14 students were trained for one hour in each day of a week on four of expository discourse types: *collection of description*, *causation, problem/solution*, and *comparison* (Meyer, 1975; 1985). During the training session, they were guided to read examples of naturally-

occurring passages and were asked to appreciate the different types of text structures. The 11 students in the control group were organized to focus on linguistic features, such as grammar and vocabulary, as well as the content of the text by answering questions and discussing. The results indicated that training on the top-level organization patterns significantly increased the amount of information that the students recalled. Tang (1992) explored the effect of teacher-provided graphic representation of text structures on 45 seventh grade ESL students. The results showed that the graphs facilitated reading comprehension and immediate recall.

Researchers have employed various methods to instruct text structures to enhance readers' awareness of text organization mechanisms and discourse knowledge, such as graphic organizers, text maps, outline grids, tree diagrams, higher order summaries, and identification of top-level organization patterns. It might be justified to infer that structure instructions lead to readers' enhancement of their discourse knowledge, which further contributes to their improvement in reading recall and comprehension.

Compared with the diversity of approaches to text structure instruction, measurements of discourse knowledge are scarce. In L1 reading research, Sanchez & Garcia (2009), aiming at exploring the influence of rhetorical competence on reading comprehension, designed two types of tasks to measure the knowledge of textual integration of the participants, 185 sixth-graders. The first task was to ask the pupils to read a passage containing anaphors (*e.g., this phenomenon*) and to find their antecedents.

The second task required the participants to read 10 passages and write the continuation to evaluate if they had grasped the global structure.

Vongpumivitch (2004) examined the nature of text structure by investigating the performance of 125 ESL students on four different tasks, an incomplete outline, open-ended questions, a graphic organizer and a summary. All these tasks were based on one passage, which is a collection of descriptions. The incomplete outline task required participants to complete a partially-blank outline with major ideas and supporting details. Seven open-ended questions asked the participants about the main idea, the major ideas of each paragraph, and the top-level structure of the text. The summary task was to assess how well the participants recognized the hierarchy of the ideas. The graphic organizer task was to ask the participants to fill in a table in their own words about the overall main idea of the passage, the major ideas and the supporting details. An additional task, thinking aloud protocol, was administered to 17 volunteers. The results suggested that the construct, *knowledge of text structure*, can be measured by the four tasks. However, there existed interaction between the types of tasks and the participants' performance.

Abeywickrama (2007) employed the rational deletion gap-fill, or the rational deletion cloze, as well as composition and summary to measure the knowledge of textual cohesion and coherence. Cohesion was defined as the linguistic features that signal connections between sentences and tie together the propositions in texts, which corresponds to Meyer's (1975, 1985) microproposition and macroproposition levels of text structure. Coherence referred to the overall discourse level unity or the global quality

of text structure, which is similar to Meyer's (1975, 1985) top-level organization. The results revealed that the rational deletion cloze is a valid approach to measure discourse knowledge.

Given that a number of studies have employed Meyer's classification of top-level organization of expository texts (e.g., Bartlett, 1978; Carrell, 1985; Mayer & Poon, 2001) to enhance readers' awareness of discourse mechanisms, it is rational to draw on the method of top-level structure recognition to tap discourse knowledge level.

Based on the above overview, two approaches were adopted in this study to assess the participants' discourse knowledge: the rational deletion cloze and recognition of top-level organization. Abeywickrama's (2007) cloze passage was used for the first task (Appendix I, Task A). First, English proficiency of the participants in her study and in mine are of similar level. They are university students and have learned English for about eight years. Second, the task has high reliability, with a generalizability coefficient of .87. A permission letter from Abeywickrama has been obtained.

Another task used to measure discourse knowledge was recognition of top-level organization. Six passages have been selected, and each one is followed by a multiple-choice question about the overall structure (Appendix I, Task B).

3.2.6 The instrument for measuring metacognitive reading strategy use

Influential L2 reading models have all embraced strategic skills as an important component of reading ability (e.g. Grabe, 2009; Koda, 2005; Grabe & Stoller, 2002; Urquhart & Weir, 1998; van Dijk & Kintsch , 1983). The term "strategies" is defined as

"actions selected deliberately to achieve particular goals" (Paris et al., 1991, p. 610). Reading strategies can be characterized by three core elements: deliberate, goal/problem-oriented, and reader-initiated/controlled (Koda, 2005).

Scholars often contrast strategies with skills. According to Alexander et al. (1998) strategies and skills differ in various ways although they are both procedural knowledge. Skills are routinized and automatic, but strategies are deliberate, intentional, and conscious. Processing skills takes minimal expenditure of cognitive effort, but strategies are effortful and take much of cognitive resources.

Classifications of strategies

Scholars categorize strategies in a variety of ways. Paris et al. (1991) grouped strategies based on time of use: *before, during, and after*. Some researchers grouped them according their functions (Anderson et al. 1991): supervising, supporting, paraphrasing, establishing coherence in a text, and test taking. Some researchers classified them as local/global or bottom-up/top-down strategies (Abbott, 2006; Barnett, 1988; Carrell, 1989; Plakans, 2009). Local or bottom-up strategies center on word level meaning, sentence structure, and textual details. Top-down, or global, strategies focus on main ideas, discourse organization, and the use of background knowledge. Chamot & O'Malley (1994) categorized strategies into three groups: cognitive, metacognitive, and social and affective strategies. Phakiti (2003) and Purpura (1998) also used this way of categorization. Cognitive strategies are used for accomplishing a specific cognitive task during reading, such as inference and word part analysis. Metacognitive strategies are

used to regulate cognitive processing as in comprehension monitoring and repair, while social and affective strategies are used when interacting cooperatively with others during reading, *e.g.* seeking outside assistance.

Cognitive strategies and metacognitive strategies

The term "metacognition" was first used by Flavell (1979), to refer to the part of one's acquired world knowledge that has to do with cognitive matters. It is cognition of cognition, or thinking about thinking. Baker & Brown (1984) recognized two dimensions of metacognition: knowledge about cognition and regulation of cognition. These two dimensions can be also termed metacognitive awareness and metacognitive strategies. For the first dimension, a reader with metacognitive awareness knows about his cognitive resources, his strengths and limitations as a reader. For the second dimension, a reader uses metacognitive strategies to regulate the cognitive process of reading. Metacognitive strategies are distinguished from cognitive strategies. Cognitive strategies are invoked to make cognitive progress, such as using suffix to guess the meaning of an unfamiliar word, while metacognitive strategies are used to monitor cognitive progress, such as self checking comprehension when reading. L1 reading researchers, such as Brown (1980) and Baker & Brown (1984) grouped metacognitive reading strategies into six clusters:

- Clarifying the purpose of reading
- Checking whether comprehension is occurring
- Self-questioning to determine whether goals are being achieved
- Identifying the important aspects of a message

- Focusing attention on the major content rather than trivia
- Taking corrective actions when failures in comprehension are detected.

In one of the earliest L2 strategy studies, Hosenfeld (1977) generalized four types of strategies from the 210 students' think-aloud data: 1) keeping the meaning of the passage in mind during reading; 2) reading in "broad phrase"; 3) skipping unimportant words; 4) having a positive self-concept as a reader. Hosenfeld did not make the difference between cognitive and metacognitive strategies.

Some studies examined both L1 and L2 readers, with intention to compare whether they employ different strategies in L1 and L2 reading. Block (1986) categorized 15 types of cognitive and metacognitive strategies from L1 and L2 readers. Among them three kinds of metacognitive strategies were identified: 1) comment on behavior and process; 2) monitor comprehension; and 3) correct behavior.

Some researchers believe that L1 and L2 readers use different strategies and L2 learners with different L1 background also utilize different strategies (e.g., Abbott, 2006). However, differences are only found in cognitive strategy use not in metacognitive strategy use. Abbott (2006) found the differences between Chinese ESL and Arabic ESL learners' strategy use. Kwon (2010) found that metacognitive strategy use varies little across L1 or L2 reading.

Strategic competence and reading ability

Research has consistently found that strategic competence is related to reading proficiency. Brantmeier (2002) reviewed 13 studies in this research line. An early study,

Hosenfeld (1977), using think-aloud reports, found that successful readers kept the meaning of a passage in mind, whereas poor readers focused on solving unknown words or phrases. In the same vein, by comparing oral reports from proficient and non-proficient readers of English, Block (1992) found that more proficient readers relied more on meaning-based global strategies, while less proficient readers used more word-based local strategies. A recent study by Plakans (2009) explored reading strategies in integrated L2 writing tasks. The study revealed that higher scoring writers used more global strategies than lower scoring writers.

Some studies take L1 and L2 background into consideration. Carrell (1989) examined the strategy use of 75 native English speakers learning Spanish and 45 native speakers of Spanish in ESL courses and found that lower proficiency levels of Spanish learners used more bottom-up processing strategies, whereas advanced ESL group used top-down strategies.

Anderson (1991), using a 47-item strategy inventory to categorized 28 Spanish speaking ESL learners' think-aloud records, found that students who used more frequently strategies comprehended better but no significant relationship existed between the amount of unique strategies and comprehension.

Schoonen et al. (1998) explored the relative contribution of metacognitive and language-specific knowledge to 685 6^{th} to 8^{th} graders' native and foreign language reading comprehension with a self-designed instrument of four domains of metacognitive knowledge, namely, assessment of oneself as a reader, knowledge of reading goals and

comprehension criteria, knowledge of text characteristics, and knowledge of reading strategies. The study revealed that for older students, metacognitive knowledge appeared to play a significant role in both native and foreign language reading comprehension.

Van Gelderen et al. (2004), with a similar research purpose to that of Schoonen et al. (1998), compared the relative contribution of metacognitive skills, linguistic knowledge, and processing speed to reading comprehension for 397 8^{th} to 10^{th} graders' first- and second-language. The results showed that metacognitive knowledge is an important predictor for both L1 and L2 reading comprehension.

Phakiti (2008) examined that relationship of EFL test-takers' long-term strategic knowledge and actual strategy use to second language reading test performance over time. The study revealed that metacognitive strategy use directly affected cognitive strategy use, and long-term cognitive strategy use directly affected language test performance to varying degrees.

Based on the above review of strategy theories and empirical studies, Phakiti's (2008) cognitive and metacognitive strategy questionnaire was revised for the present study. A survey of 22 items was designed (Appendix J). These 22 items cover seven domains of cognitive and metacognitive strategies, i.e., comprehension, retrieving, planning, monitoring, regulating, and comprehension repairing and evaluating. As shown in Table 3.1, each dimension of strategy is measured by one to give items.

Table 3.1 Strategy dimensions and their measuring items

Strategy type	Domains	Number of items	Items
Cognitive strategies	Comprehending	5	4, 5, 6, 7, 8
	Retrieving	4	9, 10, 11, 12
Metacognitive strategies	Planning	3	1, 2, 3
	Monitoring	6	13, 14, 15,16,17
	Regulating	2	18, 19
	Repairing	1	20
	Evaluating	2	21, 22

3.3 DATA COLLECTION

The data for the present study were collected in two stages. The first stage gathered the information of participants' skills of the six components that were used to model reading ability. This stage began five weeks after the participants took the CET and lasted for 30 consecutive days. The second stage focused on the collection of participants' scores of the CET reading section.

3.3.1 Administration of the six instruments for measuring reading components

The test was administered in an office of the Foreign Language School of the aforementioned university, between 8:00am to 6:00pm, for about a month. When prospective participants arrived at the office, the investigator or one of the two research assistants explained the goals, procedures, benefits, and risks of the study. Then they provide the Consent Form (Appendix K) for the students to read. Given that the students

have studied English for about eight years, the Consent Form was not translated into Chinese. However, when students required clarification, the investigator explained in Chinese. When a student decided to participate, they were asked to sign the Consent Form.

Then the participants were asked to present their CET admission cards, and the numbers of the cards were recorded for the collection of their scores in the CET reading section. After that, the participants were required to complete a questionnaire about their personal information (Appendix C). Then, they began to do a DXDX programmed word recognition efficiency test on an IBM X200 laptop as well as a working memory test which involved writing 28 words attached to each sentence on a sheet. Finally, they were required to complete a questionnaire about strategy use in English reading, as well as three paper-pencil instruments on vocabulary knowledge, syntactic knowledge, and discourse knowledge. These four parts were combined together. The type of activities, length of time, number of testing items, and administration modes are listed in Table 3.2.

Table 3.2 Process of data collection

Phase	**Length of time**
Part I: Preparation	
1. Introduction & consent form signing	5 mins
2. Personal background questionnaire	5 mins
Part II: Administration of the six measurements	

1. Word recognition	5 mins
2. Working memory	10 mins
3. Metacognitive reading skills	10 mins
4. Semantic knowledge	15 mins
5. Syntactic knowledge	20 mins
6. Discourse knowledge	20 mins
Total	**90 mins**

3.3.2 Collection of the scores in the CET reading section

Three months after the administration of the CET, the researcher presented to the Foreign Language School of the aforementioned university a list of the numbers of the participants' CET admission cards and requested their scores in the CET reading section. The scores were released to the researcher a week later.

3.4 DATA ANALYSIS

Data analysis underwent three stages: the item level, the variable level, and the structural equation modeling level. At the item level, tasks in word recognition, working memory, and vocabulary generated more than one type of raw data, and their respective data analysis methods are illustrated below. There is only type of scores for syntactic and discourse knowledge, as well as metacognitive strategy use in reading. The scoring methods for these three tasks were straight forward, i.e., the summation of item scores.

3.4.1 Software

Three software programs were utilized for the data analysis process of the present study. First, DMDX Analyze 5.1.0 was used for word recognition and working memory tasks. It generated mean reaction time and the response correct rate for each participant in data summary file (.das).

Second, SPSS 17.0 was used for descriptive statics, reliability analyses, and correlation matrices. Third, Mplus version 6.1 was employed for confirmatory factor analysis and structural equation modeling.

3.4.2 Number of items for each participant

The data analysis was based on the raw scores of the 323 items for each participant, as well as their scores in the CET reading section. As shown in Table 3.3, word recognition has 25 items, and working memory has 28 items. Three sets of tests were used to measure semantic knowledge, and each set has 60 items (40 real words and 20 pseudowords). The task used to measure syntactic knowledge has 32 items. Two tasks were employed to measure discourse knowledge. The first task includes 30 items and the second task involved 6 mini passages. The questionnaire about metacognitive strategy use in reading includes 22 items. In sum, the total item during the first stage of data collection is 323.

In the second stage, the participants' scores in the CET reading section was collected. Although the CET reading section is composed of three types of reading tasks:

fast reading, reading in depth, and banked cloze, only the total score was accessible to the researcher.

Table 3.3 Number of items and types of raw scores for each participant

Variable	**Number of items**	**Types of data collected**
Word recognition	25	1. Reaction time to each item
		2. Correct or wrong answer
Working memory	28	1. Reaction time to each sentence
		2. Judgment to the statement of the sentence
		3. Recall of the attached word
Semantic knowledge	180	1. Yes or No to a real word
		2. Yes or No to a pseudoword
Syntactic knowledge	32	Correct or wrong answer
Discourse knowledge	36	Correct or wrong answer
Strategies	22	Frequency level of strategy use
CET reading		Total score of the CET reading section
Total	**323**	

3.4.3 Scoring methods for the word recognition task

The software DMDX Analyze 5.1.0 was used to analyze participants' responses to the 25 items. An input specification file was written in Notepad and saved in a specification file (.spc). This file specifies how the data should be treated. Analyze 5.1.0 generated two files: item summary file (.ism) and data summary file (.das). The data

summary provided each participant's response accuracy and mean of reaction time. Response accuracy is the percentage of correct answers. Reaction time is defined as the length of time between the appearance of a stimulus and a participant's response to it. Outliers are set as any reaction time outside 2.0 standard deviations of the same participant's mean reaction time. Any reaction time shorter than 200 milliseconds was excluded from analysis, because such a short reaction time implies the possibility of guessing rather than actual cognitive processing.

3.4.4 Scoring methods for the working memory task

Three types of data were collected, i.e., recall, correct judgment rate, and reaction time. The recall task was scored by judging the sentence level the participants were able to attain — from zero to five. They had to successfully complete two sets at each level to get the full scores. If the participants were able to write out the words correctly and in the correct order for only one set, they were awarded a score between the level of the task and one level lower. For example, if the participant was only able to write out the three words in one three-sentence level set, then his score was not three but 2.5.

Similar to the procedure of word recognition, an input specification file was written in Notepad and saved in the specification file format (.spc). Correct judgment rate and reaction time were generated from the software DMDX Analyze 5.1.0 and presented in the data summary file (.das). Correct judgment rate is the percentage rate of correctly judging the statement of the 28 sentences. Reaction time is the mean of reaction times to the correctly judged sentences. The reaction time to the falsely judged sentences was

excluded from the calculation. Items for which a participant's reaction time was beyond 2.0 standard deviations of the mean reaction time of a participant were excluded. Items for which reaction time was shorter than 300 milliseconds were excluded from data analysis because this indicated that the reaction was more of guess than an actual judgment.

3.4.5 Scoring methods for the semantic knowledge measurement task

The question of how to yield a meaningful score is a challenge to the Yes/No test (Huibregtse et al. 2002). As shown in Figure 3.1, response to each item can be categorized into four types, i.e., hit, false alarm, miss, and correct rejection. If a particular item is a real word, and the participant gives a positive answer to it indicating he knows the basic meaning of the word, the answer is categorized as hit. If a particular item is a pseudoword, and the participants gives positive answer to it, his answer is termed as false alarm. If the participant does not check a real word, his answer to that word is categorized as miss. Finally, if the participant does not check a pseudo word, his answer belongs to correct rejection.

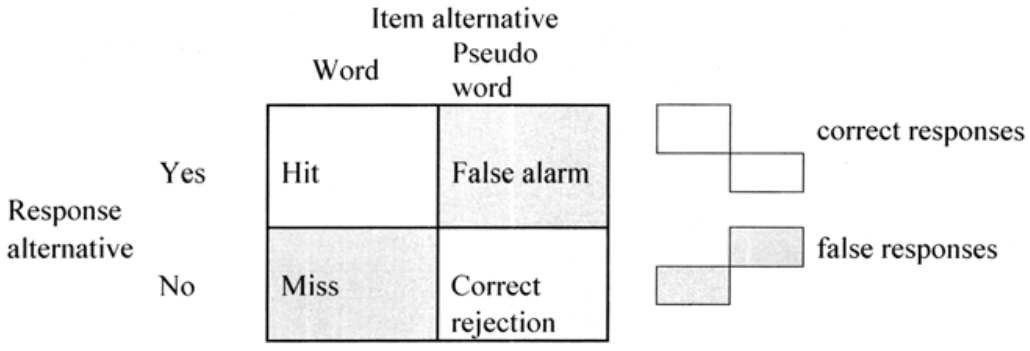


Figure 3.1 The item-response matrix of the Yes/No test

Currently five scoring methods are available (cf. Beeckmans et al. 2001; Huibregtse et al. 2002; Mochida & Harrington, 2006). They are 1) the number of correct responses; 2) proportion of hits minus the false alarm rate; 3) correction of guessing; 4) Meara's Δm; and 5) Huibregtse et al.'s (2002) new index based on Signal Detection Theory. The formulas of the methods are as follows.

1) The number of correct responses

   Score = N (h) + N (f)

   Where N (h) = number of total hits, N (f) = number of total false alarms

2) Proportion of hits minus the false alarm rate

   P (h) = $h - f$

   where P(h) = true hit rate, h = observed hit rate, f = observed false alarm rate

3) Correction of guessing

   P (h) = $\frac{h - f}{1 - f}$

   where P(h) = true hit rate, h = observed hit rate, f = observed false alarm rate.

4) Meara's Δm

   Δm = $\frac{(h - f)}{(1 - f)} - \frac{f}{h}$

   Where h = observed hit rate (number of hits divided by the total number of true words)

   f = observed false alarm rate (number of false alarms divided by the total number of pseudowords)

5) Index based on Signal Detection Theory

$$I_{SDT} = 1 - \frac{4h(1-f) - 2(h-f)(1+h-f)}{4h(1-f) - (h-f)(1+h-f)}$$

Where h = observed hit rate, f = observed false alarm rate

In the current study method 5), the index based on Signal Detection Theory, was employed. First, it is the only measure that corrects both guessing and participants' response style (Huibregtse et al., 2002). Three other methods, 2), 3) and 4) only adjust scores for guessing. Second, Meara's Δm tends to underestimate vocabulary knowledge and correction for guessing yields overestimations (Huibregtse et al. 2002: 240). Third, although the simplest way to score a Yes/No test would be method 1), to count the number of correct responses, which is the sum of the hits and the correct rejections of pseudowords, it is not appropriate to treat two types of correct responses equivalently. The first type estimates word knowledge but the second type of correct responses means either the test taker does not know the word or he knows it is not a real word. Neither is an indication of vocabulary knowledge. In sum, index based on Signal Detection Theory is used to analyze participants' response to the 180 items in the three sets of vocabulary test.

3.4.6 Scoring methods for syntactic knowledge measurement tasks

The task was scored by awarding one point to each item that was correctly answered and zero for those incorrect answers. The highest possible score is 32, and lowest is zero.

3.4.7 Scoring methods for discourse knowledge measurement tasks

For Task A, one point was awarded to each correct answer and zero to each incorrect answer. Task B was scored the same way as that of Task A. These two tasks tap different dimensions of discourse knowledge. Task A focuses more on cohesion within and between sentences, while Task B taps only top-level organization recognition. Therefore, the summation of these two tasks is used as the index of participants' discourse knowledge.

3.4.8 Scoring methods for metacognitive strategy use measurement tasks

After each of the 22 statement of strategy use in reading there are six words, i.e., *Never, Rarely, Sometimes, Often, Usually, and Always*, indicating the different frequency of strategy use. *Never* is coded zero, *Rarely* 1 point, *Sometimes* 2 points, *Often* 3 points, *Usually* 4 points, and *Always* 5 points. The summation of the scores on the 22 items is the total score indicating the participant's metacognitive strategy use in reading.

3.5 VARIABLE INDEX

Among the six variables used to model the latent variable of reading ability, four variables were indexed by one type of score. However, the word recognition variable has two indices, correct response rate and reaction time. The working memory variable was indexed by correct response time, reaction time, and recall.

Researchers have employed various approaches to yield the index for the variable with more than one indicator. Some computed a composite score (e.g., Stanovich & West, 1989; Nassaji & Geva, 1999). These researchers first standardized the number of error responses and the time taken to perform the task. The two resulting *z* scores were

then averaged to yield a single composite score. Others (e.g., Wang & Koda, 2007) examined accuracies and reaction times separately. It is customary to use only reaction time as the measurement of processing speed in word recognition research (Juffs, 2001). Similarly, in working memory studies, recall has been generally used as the index of working memory capacity (e.g., Conway et al., 2002). However, Juffs (2011) questioned the practice of neglecting the variation in correct response rates. In the present study, two models were compared. The first model incorporated word recognition reaction times, word recognition correct response rates, working memory reaction times, and working memory correct response rates, recall, as well as the other four variables with only one index, which is called nine-observed-variable model. The second model excluded three indices, i.e., word recognition correct response rates, working memory reaction times, and working memory correct response rates, which is a common practice in word recognition and working memory research. The second is labeled as six-observed-variable model. The two confirmatory models were analyzed and the model fit indices were compared, and the better one was chosen as the baseline model for structural equation model analysis.

3.6 STATISTICAL PROCEDURES

Three major steps were conducted for statistical analysis. The first step was at the item level. The second step involved confirmatory factor analysis of reading ability. The third step was the structural model analysis including the scores in the CET reading section. These steps are described in detail in the sections that follow.

3.6.1 Item-level data analysis

Item-level descriptive statistics computation comprised the initial step of the statistical analysis. Participants' responses to each item were first inputted to SPSS. Then the mean of each measure, the mean for each participant, and the standard deviations were computed. Next, the normality of these data was examined by associated skewness and the kurtosis values.

Based on participants' responses to each item, the reliability of the instruments was calculated by SPSS in terms of Cronbach's alpha, which is a common measure used to check instruments' internal consistency. This measure can be viewed as an extension of the Kuder-Richardson Formula 20 (KR-20). While KR-20 treats half of an instrument as parallel to the other half, Cronbach's alpha views each item as parallel to any other item and each item measures the same construct. Thus, Cronbach's alpha generally increases as the intercorrelations among test items increase.

3.6.2 Confirmatory factor analysis

Confirmatory factor analysis was conducted in four steps. First, the nine indices, i.e., word recognition reaction times, word recognition correct response rates, working memory reaction times, and working memory correct response rates, recall, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy, were treated as indicators of the latent variable of reading ability. The confirmatory model was analyzed and respecified. Second, the six-observed-variable one factor model was analyzed and respecified. Third, two higher order confirmatory models were analyzed.

Finally, these four models were compared and the one with the best model fit indices were chosen as the baseline model for the structural equation model.

3.6.3 Structural equation modeling analysis

The final step involved the structural equation modeling (SEM) analysis of the relationship between reading ability and test-takers' performance on the CET reading section. SEM gleans the functions of multiple regression, path analysis, and factor analysis to test hypothesized interrelationships among a set of variables (Bentler, 1995). The interrelationship is generally sketched out with path diagrams. Various sets of symbols are used for path diagrams, and Figure 3.2 illustrates the symbols used in this study.

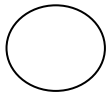
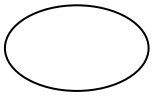
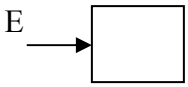
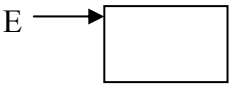
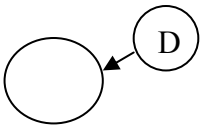
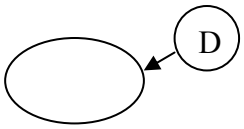
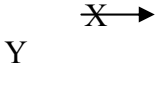
Diagram		Notation
		A latent variable (construct, or factor) not measured directly
		An observed, measured variable
		Error or residual in a measured variable not explained
		Disturbance or residual in a latent variable not explained
 Y		Theory implies X might influence Y, but not vice versa.

Figure 3.2 Path diagram symbols for structural equation models

3.6.4 Model evaluation

All confirmatory models and structural models were judged by the joint criteria suggested by Hu & Bentler (1999). The combination of these rules can minimize the possibility of rejecting the correct model and retaining the incorrect one. Table 3.4 shows the fit indices and their respective criterion.

Table 3.4 Fit indices and statistical criteria

Fit Index	Criterion
Chi-square (χ^2) test	Non-significant at the level of .05
Bentler comparative fit index (CFI)	Larger or equal to .96
Root mean square error of approximation (RMSEA)	Smaller than .06
Standardized root mean square residual (SRMR)	Smaller than .10

3.6.5 Model modification and respecification

If model fix indices of the confirmatory models of reading ability were poor they were respecified by referring to the statistics generated by the Lagrange Multiplier (LM) Test. The model with the best model fit indices served as the baseline model to create the structural model. The scores in the CET reading section were incorporated into the structural model in order to explore the relationship between actual reading ability and the scores in the CET reading section.

In summary, this chapter has presented the methods of participant recruitment and the design of the six instruments used to measure the components for modeling reading

ability. Additionally this chapter includes a discussion of the methods of data collection and data analysis, as well as the statistical procedures utilized in the study.

Chapter 4 Results

This chapter will first report the descriptive statistics for each instrument, including the means, the standard deviations, the normality of the data, and the reliabilities of the instruments. Second, the results of the confirmatory factor analysis of reading ability will be presented. Finally, the results of the structural equation modeling analysis including the relationship between reading ability and CET scores will be documented.

4.1 Missing data treatment

Since the study was conducted individually, each participant completed all sections of the measurement but unperformed items sometimes occurred in the syntactic knowledge, discourse knowledge, and metacognitive reading strategies instruments. Unperformed items in syntactic and discourse knowledge measurements were scored zero. Regarding metacognitive strategy use, nine participants missed the last item of the instrument, which was probably resulted from the circumstance that the last item was printed on a new page and was close to the instructions of the following instrument, and they probably accidentally missed the last item. Therefore, these missing items were coded by the mean item scores of the individual participants.

4.2 Participants

A total of 181 undergraduate students from a prestigious comprehensive university in central China participated in this study. The participants were in the first or

second year of their tertiary education and were studying to obtain their bachelor degrees in a variety of fields, such as mathematics, chemistry, biology, economics, engineering, mapping, medicine, etc. Of the 181 participants, 17 attended the CET before December 18, 2010, and these participants were therefore excluded from data analysis because their scores on the CET reading section were not available. Thus, the actual number of participants for data analysis was 164 for the present study, which represents 90.61% of the total number.

Among the 164 participants, 67 (40.9%) were male and 97 (59.1%) were female. The youngest was aged 18, and the oldest was 24. Participants had an average age of 19.79 years. They had studied English as a foreign language for 8.60 years on average, ranging from 6 to 17 years. They began to learn English at 11.43 years old on average with a range from 3 to 14 years old. They had 4 hours of in-class instruction per week during the course of the normal semester. They studied or practiced English outside of the English class for an average of 3.76 hours per week with a range from 0 to 14 hours.

4.3 DESCRIPTIVE STATISTICS OF THE SIX INSTRUMENTS

This section reports the descriptive statistics of the six measurements: word recognition, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy use, along with their reliabilities.

4.3.1 Descriptive statistics of the word recognition measurement

Word recognition was measured by a pseudowords identification task. The entire task consisted of 25 pairs of pseudowords (e.g. *kake/dake, filst/ferst, thair/theer*) (Appendix E). The participants were required to indicate which letter string sounded like a real English word. The total number of observations for the word recognition measurement was 4096. Four responses were shorter than the cutoff value of 300 milliseconds and were excluded from the data analysis. A total of 68 (1.67%) responses were beyond 2.0 standard deviations of individual participants' mean reaction time and were modified. Reaction times beyond two positive standard deviations of individual participants were replaced by the individual participant's mean plus or minus two standard deviations of his or her own mean of reaction times. The data analysis was conducted by DMDX Analyze 5.1.0 with the specification file (Appendix H). As shown in Table 4.1, the mean of mean reaction time was 2100.91 milliseconds, and correct response rate was 70.27%. The skewness of mean reaction time and correct response rate were -.44 and -.51, respectively. The values of kurtosis for mean reaction time and correct response rate were .23 and .12, respectively. The absolute value of the skewness and kurtosis for mean reaction time and correct response rate fell below three, which implied univariate distribution normality (Kline, 2005).

Table 4.1 Descriptive statistics of word recognition measurement

Indicator	Min	Max	Mean	SD	Skewness	Kurtosis
Mean reaction time	1128.00	2750.00	2100.91	275.80	-.44	.23
Correct response rate	32.00	96.00	70.27	13.55	-.51	.12

Notes: Number of participants N = 164; Number of items k = 25.

4.3.2 Reliability of the word recognition measurement

In order to examine the reliability of the word recognition measurement, each of the 164 participants' responses to the 25 items were input to SPSS. Correct responses were coded 1, and incorrect answers were coded 0. Cronbach's alpha was used as the index of reliability. The alpha value for the word recognition measurement was .63.

4.3.3 Descriptive statistics for the working memory measurement

Working memory was measured by a silent sentence reading span test. The participants were required to read silently eight sets of sentences and to write down the unrelated word attached to each sentence at the end of each set (Appendix F). The total number of observations was 4590. Response times shorter than 300 milliseconds were viewed as guessing rather than normal processing. Two responses were shorter than the cutoff value of 300 milliseconds and were excluded from the data analysis. Response times beyond 2.0 standard deviations of individual participants' mean reaction time were treated as outliers and were modified. A total of 168 (3.7%) responses were modified. Reaction times beyond two positive standard deviations of individual participants were

treated as outliers and were replaced by the specific participant's mean plus two standard deviations of his own. Reaction times beyond two negative standard deviations were replaced by the specific participant's mean minus two standard deviations of his own. The data analysis was conducted by DMDX Analyze 5.1.0 with the specification file (Appendix I). As shown in Table 4.2, the mean of mean reaction time was 3093.91 milliseconds, and correct response rate was 78.61. The mean of recall was 3.63. The absolute values of the skewness and kurtosis statistics of the three indices of working memory all fell below three, which implies that the data were normally distributed.

Table 4.2 Descriptive statistics of working memory measurement

Indicator	Min	Max	Mean	SD	Skewness	Kurtosis
WM Mean reaction time	1808.70	5767.10	3093.91	633.47	1.17	1.97
WM Correct response rate	35.70	96.40	78.61	10.62	-.83	1.19
Recall	1.00	5.00	3.63	.93	-.43	-.04

Notes: Number of participants N = 164; Number of items k = 28.

4.3.4 Reliability of the working memory measurement

Cronbach's alpha is not appropriate for the reading span task in working memory measurement, because it measures the degree to which responses are consistent across the items within a single measure. Using Cronbach's alpha, each item is assumed to measure the same construct, and each item is parallel to any other item in the instrument. However, items in the working memory instrument in the present study are not parallel.

The items were grouped in two-, three-, four-, and five-sentence levels. Therefore, responses to items in the two-sentence level are not comparable to answers to items in the five-sentence level. Reponses to the two-sentence level items are very likely to yield higher correct rates than the responses to the items in the higher-level serials.

To address this problem, an approach used by Engle et al. (1999) was adopted. The scores on the first set of two-, three-, four-, and five-sentence levels were combined into a single score, as were the scores on the second set of four sentence levels. Therefore, the two total scores were comparable and used to compute split-half reliability. Table 4.3 presents the items for each set at four sentence levels. The total scores of 2A, 3A, 4A, and 5A and the total score of 2B, 3B, 4B, and 5B for each participant were recorded by SPSS. As shown in Table 4.4, the two parts were similar for the statistics of the mean, standard deviation, and spread of the scores. The split-half reliability of the working memory instrument was .50.

Table 4.3 Items for each set of the working memory task

Levels	**Sets**	**Items**
2-sentence	2A	1, 2;
	2B	3, 4;
3-sentence	3A	5, 6, 7;
	3B	8, 9, 10;
4-sentence	4A	11, 12, 13, 14;
	4B	15, 16, 17, 18;
5-sentence	5A	19, 20, 21, 22, 23;
	5B	24, 25, 26, 27, 28.

Table 4.4 Statistics of the split half of the working memory tasks

	Min	Max	Mean	SD	Skewness	Kurtosis
Score A	5	14	11.37	1.80	-.88	.88
Score B	4	14	10.64	1.83	-.76	.96

4.3.5 Descriptive statistics for the semantic knowledge measurement

In order to calculate the hit rate (the percentage of "yes" responses to authentic words) and false alarm rate (the percentage of "yes" responses to pseudowords), the participants' responses to each of the 180 items was recorded by SPSS. Six spreadsheets were employed. The first spreadsheet was used to record participants' responses to the 40 real words of the first set of the vocabulary test; the second was used to record participants' responses to the 20 pseudowords of the first set of the vocabulary test. The third and fourth spreadsheets were used for the real and pseudowords of the second set of the vocabulary test. The fifth and sixth were used for the real and pseudowords of the third set of the vocabulary test. The index based on the Signal Detection Theory (I_{SDT}) was calculated for each set, and then the three I_{SDT} were summed to a total $I_{SDT.}$ The value of the total I_{SDT} was used as the index for the variable of semantic knowledge. As shown in Table 4.3, the mean of the semantic knowledge measurement was 1.70 out of the highest possible value of 3.00. The standard deviation was .31. The skewness statistic was within [-3, 3]. The kurtosis statistic was 3.91, which implies that the data distribution was leptokurtic, featuring a higher peak. According to Kline (2005), the data of semantic knowledge were assumed to fulfill the assumption of univariate normality.

Table 4.5 Descriptive statistics of semantic knowledge measurement

Vocabulary	Min	Max	Mean	SD	Skewness	Kurtosis
Total I_{SDT}	.11	2.48	1.70	.31	-1.10	3.91

Notes: Number of participants N = 164; total number of items K = 180; number of items in each of the three sets k = 60.

4.3.6 Reliability of the semantic knowledge measurement

Following the practice of Mochida & Harrington (2006), the reliabilities of the authentic words test and pseudowords test were calculated separately, because correct responses to these two types of words have different implications (Huibregtse et al., 2002). The number of YES responses to real words can be used to estimate word knowledge, but correct responses to pseudowords imply that either the test-taker does not know the word, or the test-taker knows it is not a real word. Reliability ranged from .82 to .89 for the authentic words and from .62 to .76 for the pseudowords.

Table 4.6 Cronbach's alpha for the three sets of the Yes/No vocabulary test

	Authentic words	Number of Authentic words	Pseudowords	Number of Pseudowords
Set 1	.82	40	.62	20
Set 2	.88	40	.76	20
Set 3	.89	40	.73	20

Notes: Number of participants N = 164.

4.3.7 Descriptive statistics for the syntactic knowledge measurement

The syntactic knowledge measurement comprised 32 items. Each correctly answered item was awarded one point, and each incorrectly answered item received zero points. The total score was used as the index of the syntactic knowledge variable in the confirmatory model of reading ability.

Table 4.7 Descriptive statistics of syntactic knowledge measurement

Syntactic Knowledge	Min	Max	Mean	SD	Skewness	Kurtosis
Total score	13	32	27.17	3.52	-1.60	3.65

Notes: Number of participants N = 164; the total number of items k = 32.

4.3.8 Reliability of the syntactic knowledge measurement

The 164 participants' responses to the 32 individual syntactic items were recorded in SPSS. The total number of responses was 5246. Cronbach's alpha for the syntactic knowledge measurement was .74.

4.3.9 Descriptive statistics for the syntactic knowledge measurement

Two tasks were used to measure participants' syntactic knowledge: the 30-item rational deletion cloze (Appendix I, Part IV, Task A) and the recognition of the top-level organization of six mini passages (Appendix I, Part IV, Task B). Participants' responses to the 30 items of the rational deletion cloze and to the top-level organization of the six mini passages were recorded by SPSS in two spreadsheets. Table 4.6 shows the

descriptive statistics for the two tasks. The Pearson correlation between these two tasks was .41, which was significant at the .01 level.

Table 4.8 Descriptive statistics of discourse knowledge measurement

Discourse knowledge	Items	Min	Max	Mean	SD	Skewness	Kurtosis
Task A	30	0	23	10.68	4.75	-.08	.06
Task B	6	0	6	4.70	1.20	-.79	.46
Total	36	0	29	15.38	5.36	-.23	.23

Notes: Number of participants N = 164.

4.3.10 Reliability of the discourse knowledge measurement

The reliability of the two tasks was examined separately. The Cronbach's alpha of the 30-item rational deletion cloze was calculated based on the 4920 responses, which represent the 164 participants' answers to the 30 items. Each unanswered and incorrectly answered item was coded zero, and each correct answer was coded 1. The reliability of the 30-item rational deletion cloze was .80.

Based on the 984 responses, the Cronbach's alpha of the top-level organization task was .42. Given that there were only six mini passages in the task, this reliability is acceptable.

4.3.11 Descriptive statistics for the measurement of metacognitive strategy use

The measurement of metacognitive strategy use in reading consisted of 22 statements (Appendix I, Part I). After each of the 22 statement there were six words, i.e., *Never, Rarely, Sometimes, Often, Usually,* and *Always*, indicating the different frequencies of strategy use. *Never* was coded 0 points, *Rarely* 1 point, *Sometimes* 2 points, *Often* 3 points, *Usually* 4 points, and *Always* 5 points. The summation of the scores on the 22 items was the total score indicating the participant's metacognitive strategy use in reading. The lowest possible score was zero, and the highest possible score was 110. Table 4.7 lists the descriptive statistics of the measurement.

Table 4.9 Descriptive statistics of the measurement of reading strategies

Metacognitive Strategy	Items	Min	Max	Mean	SD	Skewness	Kurtosis
Questionnaire	22	39	96	67.27	12.05	.09	-.54

4.3.12 Reliability of the measurement of metacognitive strategy use in reading

Cronbach's alpha of the measurement of metacognitive strategy use in reading was calculated based on the 3608 responses from the 164 participants' answers to the strategy use frequency of the 22 statements. The above-mentioned coding method was used. The reliability of this measurement was .83.

4.4 DESCRIPTIVE STATISTICS FOR THE SCORES IN THE CET READING SECTION

Although the CET reading section comprises three parts with 10 items in each (see Chapter 1.2 of this paper), only the total score was accessible to the researcher. The highest possible score in reading is 249 out of the total CET score of 710. Table 4.8 shows the descriptive statistics for the scores in the CET reading section.

Table 4.10 Descriptive statistics for the scores in the CET reading section

	Min	Max	Mean	SD	Skewness	Kurtosis
Scores in CET reading section	75	249	203.99	33.97	-.164	2.96

4.5 INDICATORS OF WORD RECOGNITION AND WORKING MEMORY VARIABLES

Among the six variables used to predict the latent variable of reading ability, four variables were indexed by four individual indicators. However, the word recognition variable had two indicators: correct response rate and reaction time. The working memory variable was indexed by correct response time, reaction time, and recall.

In L1 research, reaction time data, irrespective of errors in responses, are normally used as the indicator of word recognition skills because errors are usually very low. As Juffs (2001) pointed out, however, error rates in L2 studies tend to be very high, and reaction times based on the correct-response observations pose serious problems about conclusions drawn from data excluding a large number of observations.

The present study followed the practice of L1 word recognition and working memory research due to lack of guidelines in L2 studies, participants with low correct response rates were excluded. First, correct response rates lower than 60% were transformed into missing data, which resulted in the exclusion of 29 scores of word recognition and 9 scores of working memory (see Table 4.11). Pairwise deletion was used for the correlation computation. As shown in Table 4.12, the first procedure resulted in mean correct response rates of 74.95% for word recognition and 79.99% for working memory.

Table 4.11 Percentile span of word recognition and working memory

Percentile range	**Number of participants in word recognition**	**Number of participants in working memory**
< 59%	29	8
60% - 69%	50	21
70% - 79%	38	62
80% - 89%	35	56
90% - 100%	12	17
Total	164	164

Table 4.12 Descriptive statistics of word recognition and working memory after data of lower correct response rates were excluded

Variable	Indicator	N	Mean	SD	Skewness	Kurtosis
Word recognition	Mean reaction time	135	2104.23	263.21	-.50	.32
	Correct response R	135	74.95	13.55	.38	-.71

Working memory	Mean reaction time	156	3095.79	640.15	1.17	1.96
	Correct response R	156	79.99	8.77	-.18	-.69
	Recall	156	3.64	.93	-.44	-.06

Because the correct response rates were still relatively low after the first procedure, neglecting variance in reaction times was not appropriate. To address this problem, some researchers have employed composite scores for variables with more than two indicators. Stanovich & West (1989) and Nassaji & Geva (1999) used the *z*-score average of error response rates and reaction times as the indicator of word recognition. Similarly, Walter (2004) utilized the z-score average of correct response rates, mean reaction times (multiplied by -1), and recall as the indicator of working memory. The present study, however, did not use composite scores as indicators because they tend to mask the distinct contributions of processing speed, processing accuracy, storage, reaction times, and correct response rates.

Therefore, after the exclusion of the reaction times lower than 60%, the second procedure involved conducting confirmatory analyses by incorporating all indicators of word recognition and working memory as observed variables of reading ability.

4.6 CONFIRMATORY FACTOR ANALYSIS WITH NINE OBSERVED VARIABLES

The development and evaluation of the confirmatory model of reading ability involved six steps as outlined by Kline (2005). First, a model was specified. Observed variables of the factor reading ability were presented and graphed. Second, model identification was examined. Third, the input, including the standard deviations of the

observed variables and their correlations, were presented and analyzed by the computer program Mplus 6.1. Fourth, the overall model fit was checked. Fifth, the model was respecified. Sixth, the parameter estimates were presented and interpreted. Last, competing models were examined and compared.

4.6.1 Model specification

As shown in Figure 4.1, the latent variable (factor) reading ability was indicated by nine observed variables, i.e., word recognition reaction times, word recognition correct response rates, working memory reaction times, working memory correct response rates, recall, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategies.

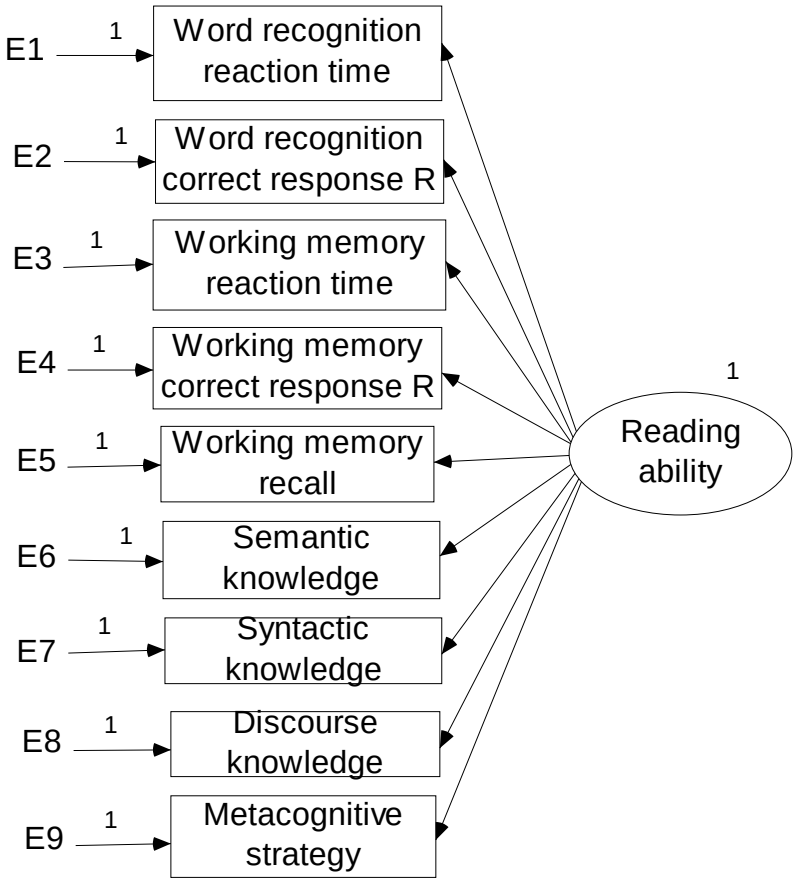


Figure 4.1 Confirmatory model for reading ability with nine observed variables

4.6.2 Model identification

There were 45 unique pieces of information in the standard deviations of the nine observed variables and their correlation matrix (9*[9+1]/2 = 45). The measurement errors were scaled to 1.0, and the variance of the factor, reading ability, was standardized and fixed to 1.0. The number of parameters requiring estimation was 18, i.e., nine variances of the nine measurement errors and nine factor loadings. Therefore, the model was over-identified because the number of unique pieces of information was greater than the

number of parameters requiring estimation. The degrees of freedom were 27 (45-18 = 27).

The sample size for the nine-observed-variable factor analysis was also examined. With 164 participants and 18 free parameters, each parameter had 9.1 participants. Therefore, the sample size was considered adequate for the factor analysis of the nine observed variables.

4.6.3 Data summary

A correlation matrix and the standard deviations of the nine observed variables were used for analysis. In order to facilitate the estimation process, the standard deviations were scaled because the biggest standard deviation was larger than the smallest one by a factor of 2000 — 640 versus .31. The former was the standard deviation for working memory (see Table 4.12), and the latter was that of semantic knowledge (see Table 4.5). Table 4.13 summarizes the scaled standard deviations and their scaling factors, and Table 4.14 presents the correlations among the nine observed variables.

Table 4.13 Scaled standard deviations of the nine observed variables

Variable	N	Original SD	Scaled SD	Scaling factor
Word recognition reaction time	135	263.21	2.63	1/100
Word recognition correct	135	9.27	9.27	1

response rate				
Working memory reaction time	156	640.15	6.40	1/100
Working memory correct response rate	156	8.77	8.77	1
Working memory recall	156	.93	9.30	10
Semantic knowledge	164	.31	3.12	10
Syntactic knowledge	164	3.52	3.52	1
Discourse knowledge	164	5.36	5.36	1
Metacognitive strategy	164	12.05	12.05	1

4.6.4 Model estimation and evaluation

The computer software Mplus 6.1 was used for model estimation. Multiple fit indices were employed to evaluate the goodness-of-fit of the model. The fit indices included 1) the chi-square (χ^2) test statistic with its level of significance, 2) the comparative fit index (CFI), 3) the root mean error of approximation (RMSEA) along with its 90% confidence interval, and 4) the standardized root mean square residual (SRMR). As shown in Table 4.15, $\chi^2(27) = 71.487$, $p < .001$, CFI = .806, SRMR = .084, and RMSEA = .100 with the 90% confidence interval .072 – .129. The chi-square test was significant at the .05 level, which implied that the null hypothesis, i.e., the model fits the data, could not be retained. The value of CFI was smaller than .95, although not too far away from the customary cut-off value. SRMR was smaller than the normally used criterion value of .10, which implied that a good fit was associated with the model. RMSEA was .10, which was greater than the pre-determined value of .06, suggesting a

poor fit. The 90% confidence interval of RMSEA was .072 to .129. The lower bound was bigger than .05, so the null hypothesis of close approximate fit was rejected. Overall, the results indicated the poor fit of the model. The indices were close to the cut-off, however, so respecification was conducted to examine whether model fit could be improved.

4.6.5 Model respecification

The Lagrange Multiplier Test (LM) was utilized for the process of model modification. The LM test is commonly used to determine whether or not model fit would be significantly improved by estimating an additional parameter.

The LM test results implied that adding eight error correlations might improve overall model fit. Among them two error correlations, working memory reaction time and working memory recall, as well as word recognition reaction time and vocabulary, were associated with the largest amount of chi-square value decrease. Considering that participants tended to recall more information if they spent less time making judgments about sentence statements, adding an error correlation between these two indicators was rational. Furthermore, these two variables corresponded to two dimensions of the construct of working memory. Adding the correlation between word recognition reaction time and vocabulary was also reasonable, given that word recognition and vocabulary both deal with translating print letter strings or characters into their meanings.

Table 4.14 Correction Matrix for the nine observed variables

	WRRT	WRCR	WMRT	WMCR	WMRC	SEMK	SYNK	DISK	METAS
Scaled SD	2.63	9.27	6.40	8.77	9.30	3.12	3.52	5.36	12.05
WRRT	1								
WRCR	-.204*0	1							
WMRT	.228**	-.146	1						
WMCR	-.223*0	.130	-.101	1					
WMRC	-.201*0	.189*	-.394**	.165*	1				
SEMK	-.038*	.268**	-.162*	.249**	.173*	1			
SYNK	-.198*	.280**	-.097	.242**	.070	.514**	1		
DISK	-.200*	.326**	-.164*	.271**	.063	.458**	.588**	1	
METAS	-.136	.151	-.163*	.115	.167*	.248**	.144	.200*	1
Reliability		.63		.50		.86	.74	.80	.83

Notes: * correlation is significant at the 0.05 level (2-tailed); ** correlation is significant at the 0.01 level (2-tailed); WRRT for word recognition reaction time; WRCR for word recognition correct response rate; WMRT for working memory reaction time; WMCR for working memory correct response rate; WMRC for working memory recall; SEMK for semantic knowledge; SYNK for syntactic knowledge; DISK for discourse knowledge; METAS for metacognitive strategy use; the reliability of semantic knowledge is the average of the reliabilities of the authentic words in the three Yes/No tests.

Table 4.15 Model fit indices for the nine-observed-variable CFA model

Index	χ^2	*df*	*p*	CFI	RMSEA	90C.I. of RMSEA	SRMR
Value	71.487	27	<.001	.806	.100	.072 - .129	.084

Two error correlations were added to the model (see Figure 4.2), and the degrees of freedom decreased from 27 to 25. The results revealed improvement of overall model fit. As shown in Table 4.16, χ^2 (25) = 31.388, p >.05, CFI = .92, SRMR = .039, and RMSEA = .100 with the 90% confidence interval .000 – .078.

Table 4.16 Fit indices for the respecified nine-observed-variable CFA model

Index	χ^2	*df*	*p*	CFI	RMSEA	90C.I. of RMSEA	SRMR
Value	31.388	25	.177	.972	.039	.000 - .078	.054

The chi-square test was not significant at the .05 level, which implied that the null hypothesis, i.e., the model fits the data, could be retained. The value of CFI was bigger than .95, indicating good fit of the model. RMSEA was .039, which was smaller than the pre-determined value of .06, suggesting good fit. The 90% confidence interval of RMSEA was .000 to .078. The lower bound was smaller than .05, so the null hypothesis of close approximate fit was retained. At .054, SRMR was smaller than the normally used criterion value of .10, which implied that a good fit was associated with the model. Overall, the results indicated the good fit of the respecified model.

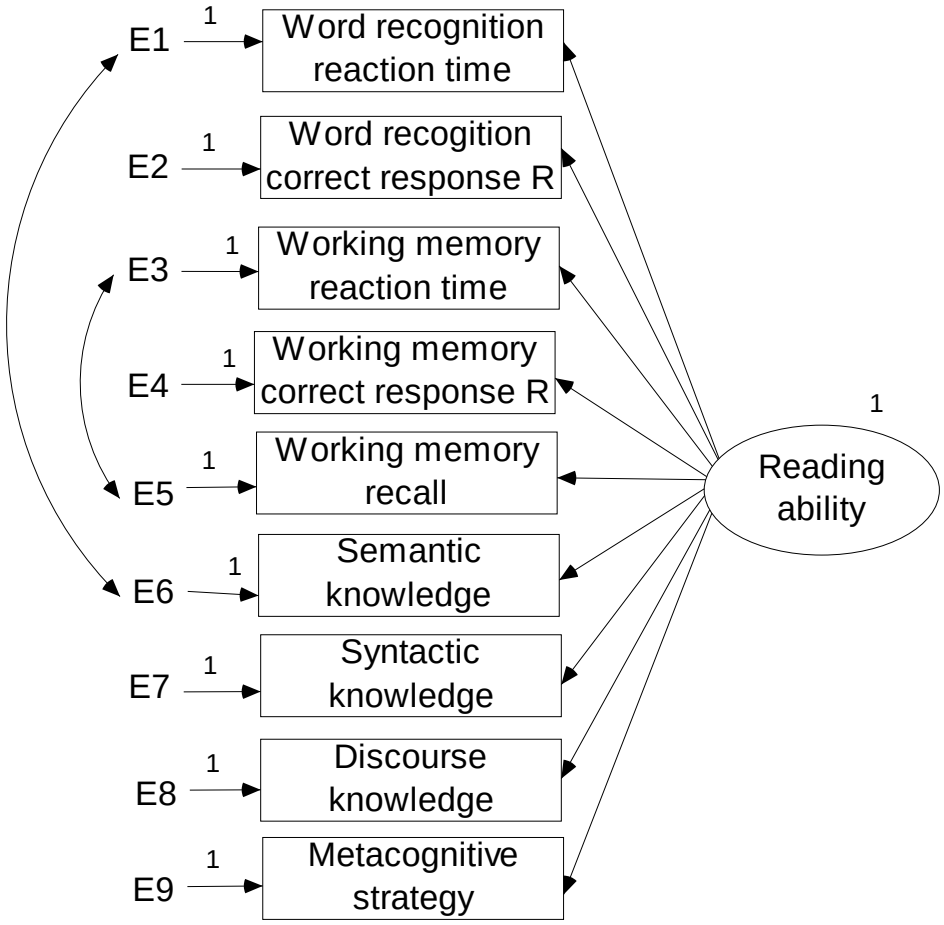


Figure 4.2 Respecified CFA model for reading ability with nine observed variables

4.6.6 Parameter estimates interpretation

First, factor loadings were reported and interpreted. Factor loadings reflect regression coefficients for the prediction of the indicators by the latent variable. In the present study, the factor loadings from reading ability to the nine indicators reflect how well reading ability predicts the observed variables, and to what degree an indicator and its latent construct are related. Whether the factor loadings were positive or negative was

examined first, and then the size and statistical significance of the factor loadings were scrutinized.

Second, the squared multiple regressions of the nine indicators were examined. The value of R^2 denotes the percentage of a given indicator's variance as explained by their latent construct. The significance of each R^2 was also reported. The information for the factor loadings and R^2 could be used to detect potential redundant indicators for the sake of model parsimony.

As shown in Figure 4.3, all the factor loadings were positive except those of word recognition reaction time and working memory sentence judgment reaction time. These two negative factor loadings were meaningful considering that readers with higher reading abilities tend to be more automatic in word recognition and therefore need less time to process the meaning of word. Similarly, readers with higher reading abilities are more efficient in sentence processing, which resulted in shorter reaction times.

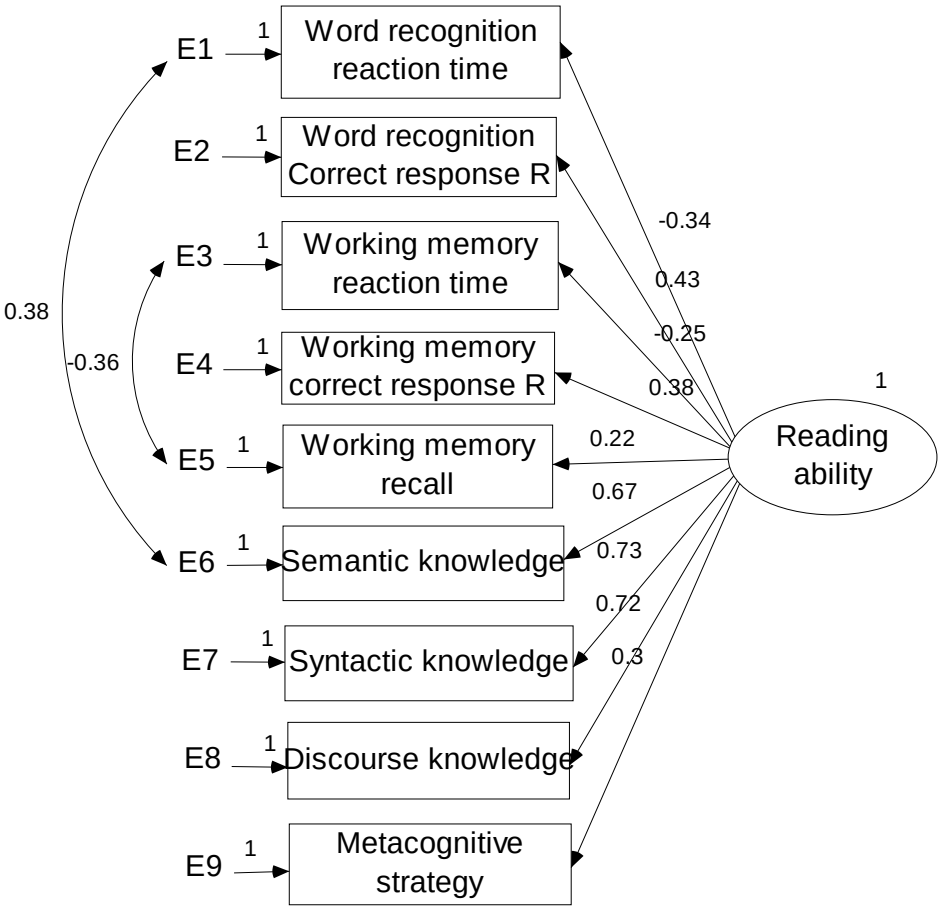


Figure 4.3 Nine-observed-variable CFA model with standardized estimates

Three factor loadings, i.e., the direct effect of reading ability on semantic knowledge, syntactic knowledge, and discourse knowledge, were approximately .70, ranging from .67 to .73. These results implied that the three variables were very good indicators of reading ability. Four factor loadings were above .30: the paths from reading ability to word recognition reaction time, word recognition correct response rate, working memory correct response rate, and metacognitive strategy. Factor loadings at these

magnitudes implied that these four variables were fair indicators of reading ability. The two lowest factoring loadings, -.25 and .22, were the paths from reading ability to working memory reaction time and to working memory recall, respectively. Such low values for the factor loadings implied that these two variables were poor indicators of reading ability. However, these two factor loadings were still significant at the .05 level. The remaining seven factor loadings were significant at the .01 level.

Table 4.17 R^2 estimates for the nine variables in the CFA model for reading ability

Observed variable	R^2 estimate	*P* value
Word recognition reaction time	.113	.042
Word recognition correct response rate	.186	.003
Working memory reaction time	.064	.125
Working memory correct response rate	.146	.012
Working memory recall	.047	.194
Semantic knowledge	.452	.000
Syntactic knowledge	.532	.000
Discourse knowledge	.519	.000
Metacognitive strategy	.092	.058

Table 4.17 presents the squared multiple regressions (R^2) and their associated significance for the nine indicators. Over 50% of the variances of semantic knowledge and discourse knowledge were accounted for by the factor of reading ability (53.2% and

51.9%, respectively). A large amount of variance of the variable of semantic knowledge was explained by the factor of reading ability. For the remaining six variables, the latent factor accounted for a very small amount of their variances, ranging from 6.4% to 18.6%. Three of the variables were marginally significant at the .05 level, i.e., word recognition reaction times, word recognition correct response rates, and working memory correct response rates. No significant amount of the variances of the other three variables, namely, working memory reaction time, working memory recall, and metacognitive strategy, was accounted for by the factor of reading ability.

As shown in Figure 4.3, the error variance correlation between working memory reaction time and recall was -.36 ($p < .001$), which implied that shorter reaction times in sentence processing resulted in better recall of the words attached to the sentences. The negative correlation between these two paths represented the competition for the limited resources of working memory between processing and memorizing.

The other error variance correlation in the respecified CFA model was the path between word recognition reaction time and semantic knowledge. The positive correlation between these variables implied that their measurement errors might come from a shared a source, such as, a similar method of measurement. Word recognition and semantic knowledge were both measured by recording participants' quick responses to stimuli. Word recognition was measured by asking participants which one in a pair of letter strings sounded like a real word, while semantic knowledge was measured by

asking participants whether they knew the basic meaning of the letter strings. Both measures involved the use of letter strings or pseudo words.

In sum, the above interpretation of the parameter estimates reveals that semantic knowledge, syntactic knowledge, and discourse were strong indicators of reading ability. Word recognition reaction time and correct response rates, working memory correct response rates, and metacognitive strategy were fair to poor indicators of reading ability. Working memory reaction time and recall were poor indicators of the factor of reading ability.

However, since the primary aim of this study is not to generate a reading ability model but to explore the relationship between reading ability and the scores in the CET reading section, this nine-observed-variable model was not further respecified.

Although the nine-observed-variable CFA model explains the data well, as indicated by the good model fit indices, other models might account for the data equally well or even better. The following section discusses three competing CFA models based on reading theories.

4.7 COMPETING CFA MODELS

Reading researchers, especially in L1, tend to use word recognition reaction time as the indicator of word recognition ability, ignoring the variation in correct response rates. This practice is reasonable if the correct response rates are high enough, for instance, 90% and above, and therefore the variance of correct response rates is small. Similarly, working memory recall is generally used as the sole indicator of working

memory ability, irrespective of working memory processing speed and quality. Juffs (2001, 2011) comments that the practices of excluding correct response rates and working memory reaction times deserves further research. The first competing model is a factor, i.e., reading ability, with six observed variables: word recognition reaction time, working memory recall, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy.

A second competing model involves Grabe & Stoller's (2002) analysis of reading ability. As shown in Figure 1.4, this higher order CFA model comprises three factors, namely, reading ability, lower-level processes and higher-level processes. Lower-level processes were indicated by word recognition, working memory, semantic knowledge, and syntactic knowledge, while higher-level processes were indicated by discourse knowledge and metacognitive strategy.

A third competing model is based on Weir's (2005) conceptualization of reading ability. As shown in Figure 1.5, this hypothesized model consists of three factors, i.e., reading ability, executive processes, and executive resources. Word recognition, working memory, and metacognitive strategy are indicators of executive processes, whereas semantic, syntactic, and discourse knowledge serve as indicators of executive resources. The following presents the confirmatory factor analyses of these three competing models.

4.7.1 Six-observed-variable CFA model for reading ability

As graphed in Figure 4.4, the latent variable reading ability was indicated by six variables, namely, word recognition reaction time, working memory recall, semantic

knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy. There were 6*(6+1)/2 = 21 unique pieces of information. All error variances were scaled to 1.0, and the variance of the factor, reading ability, was standardized and fixed to 1.0. The number of parameters requiring estimation was 12, i.e., six variances of the six measurement errors and six factor loadings. Therefore, the model was over-identified because the number of unique pieces of information was greater than the number of parameters requiring estimation. The degrees of freedom were 9 (21-12 = 9). Table 4.18 presents the input data, including a correlation matrix with the six variables' scaled standard deviations. This model was estimated by Mplus 6.1, which was the same computer program that was used for the nine-observed-variable CFA model.

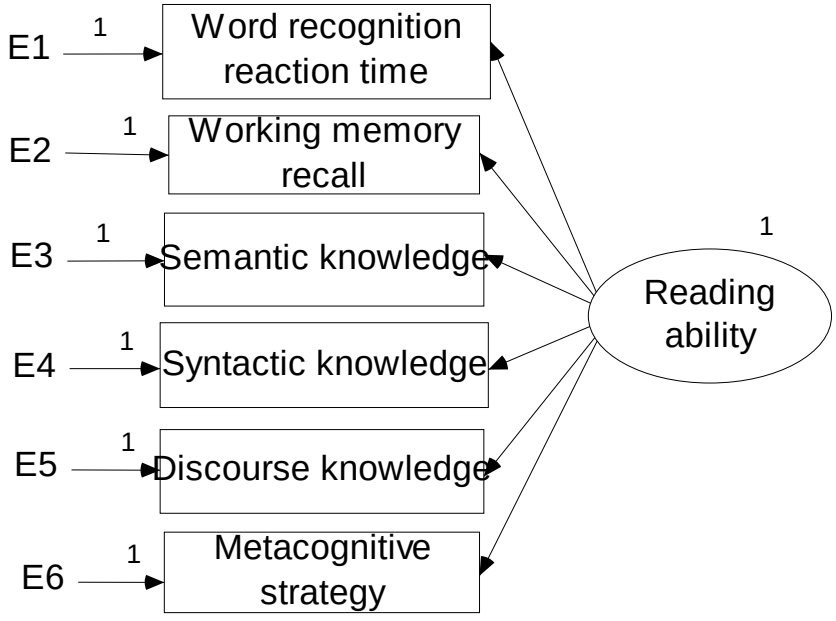


Figure 4.4 Confirmatory model for reading ability with six observed variables

Values of selected fit indices were χ^2 (9) = 29.525, p < .001, RMSEA = .118 with the 90% confidence interval .072 − .167, CFI = .871, SRMR = .071. Except for the value of SRMR, which was smaller than .10, all other indices were beyond the cut-off values. These results indicated fair to poor fit of the model to the data in Table 4.18.

Table 4.18 Correlation matrix for the six observed variables

	1	2	3	4	5	6
Scaled SD	2.63	9.35	3.12	3.52	5.36	12.05
1. Word recognition reaction time	1					
2. Working memory recall	-.201*	1				
3. Semantic knowledge	-.038	.173*	1			
4. Syntactic knowledge	-.198*	.070	.514**	1		
5. Discourse knowledge	-.200*	.063	.458**	.588**	1	
5. Metacognitive strategy	-.136	.167*	.248**	.144	.200*	1

The LM test was employed to detect whether allowing some measurement errors to covary might improve the overall model fit indices. The test results indicated that adding one of the three pairs of measurement error covariance may result in an χ^2 decrease.

These three pairs of measurement errors occurred between working memory and word recognition reaction time, semantic knowledge and word recognition reaction time, as well as metacognitive strategy and syntactic knowledge. The last pair was

meaningless, so the first two pairs of measurement error covariance were added to the model, and the degrees of freedom of the model decreased from nine to seven. The values of selected fit indices of the respecified model were χ^2 (7) = 10.658, p > .05, RMSEA = .056 with the 90% confidence interval .000 − .120, CFI = .977, SRMR = .045. These results indicated good fit of the respecified model.

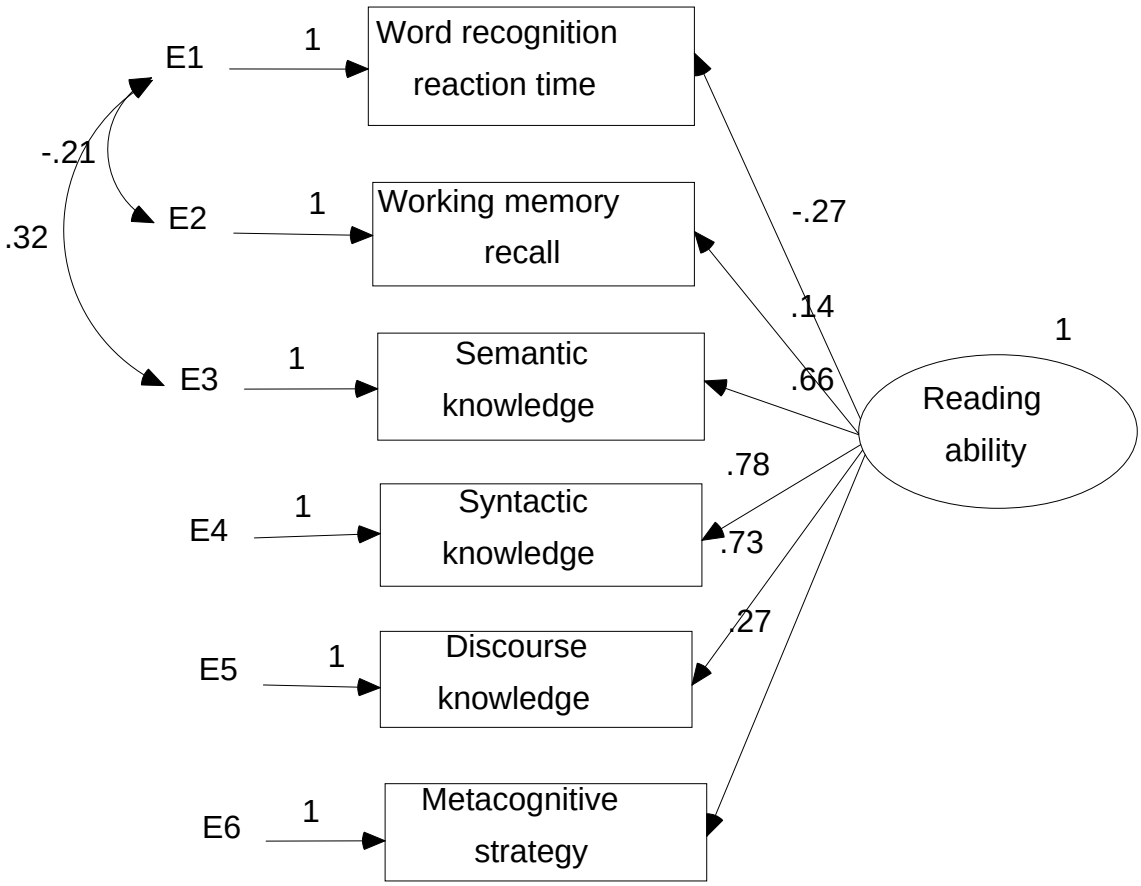


Figure 4.5 Six-observed-variable CFA model with standardized estimates

Similar to the pattern of nine-observed-variable CFA model, all of the factor loadings were positive except for the direct effect of reading ability on word recognition

reaction time. The three biggest factor loadings were still the same as those of the nine-observed-variable model, i.e., the direct effect of reading ability on semantic knowledge, syntactic knowledge, and discourse knowledge. As shown in Figure 4.5, the values ranged from .66 to .78. The R^2 for semantic knowledge, syntactic knowledge, and discourse knowledge were 43.1%, 60.7%, and 53.7%, respectively, which implied that a large amount of the variances of these three variables could be accounted for by the latent variable of reading ability. The R^2 for the other three variables were not significant at the .05 level.

Table 4.19 R^2 estimates for the six variables in the CFA model for reading ability

Observed variable	R^2 estimate	*P* value
Word recognition reaction time	.074	.111
Working memory recall	.019	.429
Semantic knowledge	.431	.000
Syntactic knowledge	.607	.000
Discourse knowledge	.537	.000
Metacognitive strategy	.075	.095

Contrary to the nine-observed-variable CFA model, the six-observed-variable model masked the underlining effect of reading ability on word recognition and working memory. In the former CFA model, the factor loadings and R^2 for word recognition correct response rate and working memory correct response rate were significant at the

.05 level. Therefore, the nine-observed-variable CFA model was favored as the baseline model for the structural model analysis.

4.7.2 High order CFA model 1 for reading ability

There were at least two possible higher order CFA models for reading ability based on L2 reading theories. The first involves Grabe & Stoller's (2002) analysis of reading ability. The second competing model is based on Weir's (2005) conceptualization of reading ability.

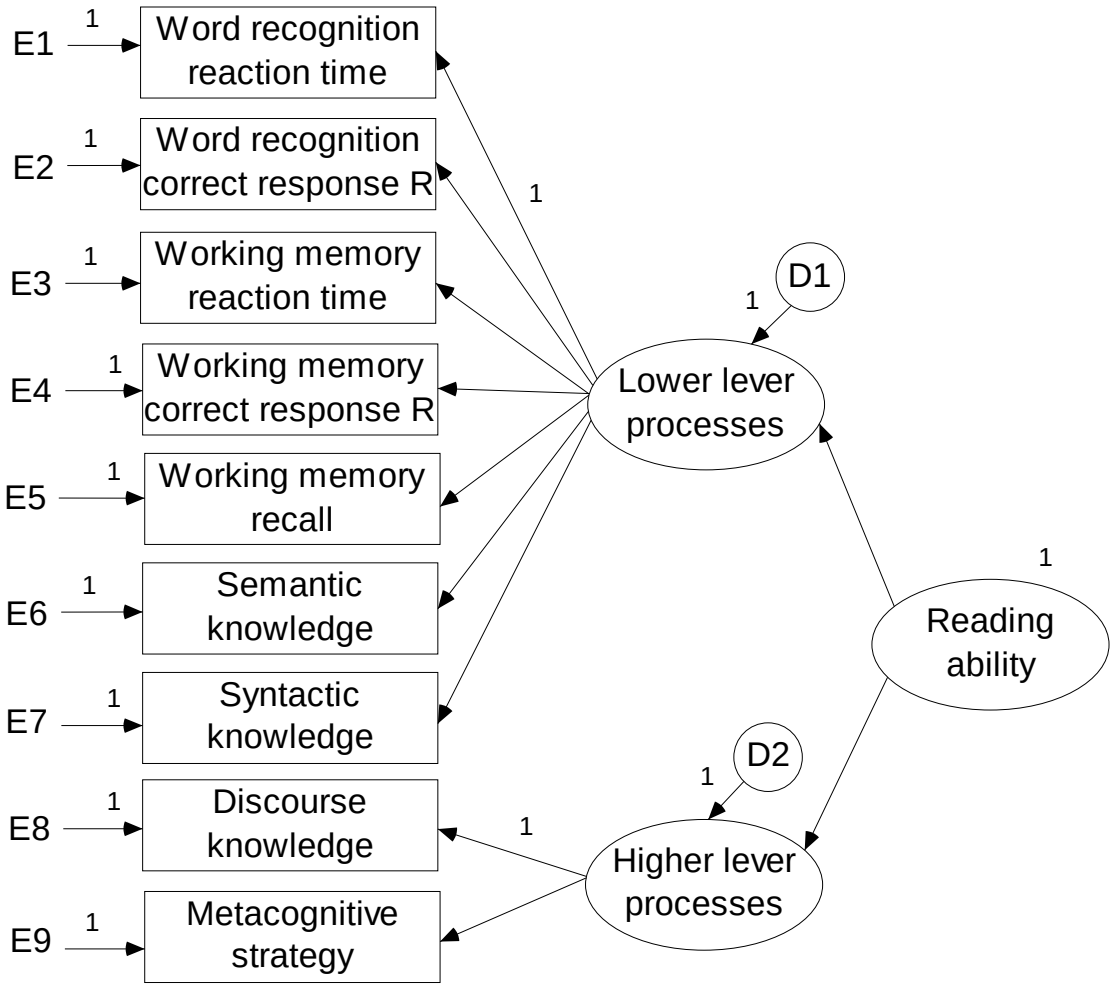


Figure 4.6 Higher order CFA model 1

As shown in Figure 4.6, the higher order CFA model 1 comprises three factors, namely, reading ability, lower-level processes, and higher-level processes. Lower-level processes were indicated by word recognition reaction time, word recognition correct response rate, working memory reaction time, correct response rate, recall, semantic knowledge, and syntactic knowledge, while higher-level processes were indicated by

discourse knowledge and metacognitive strategy. In the second order, reading ability was indicated by the lower-level and higher-level processes.

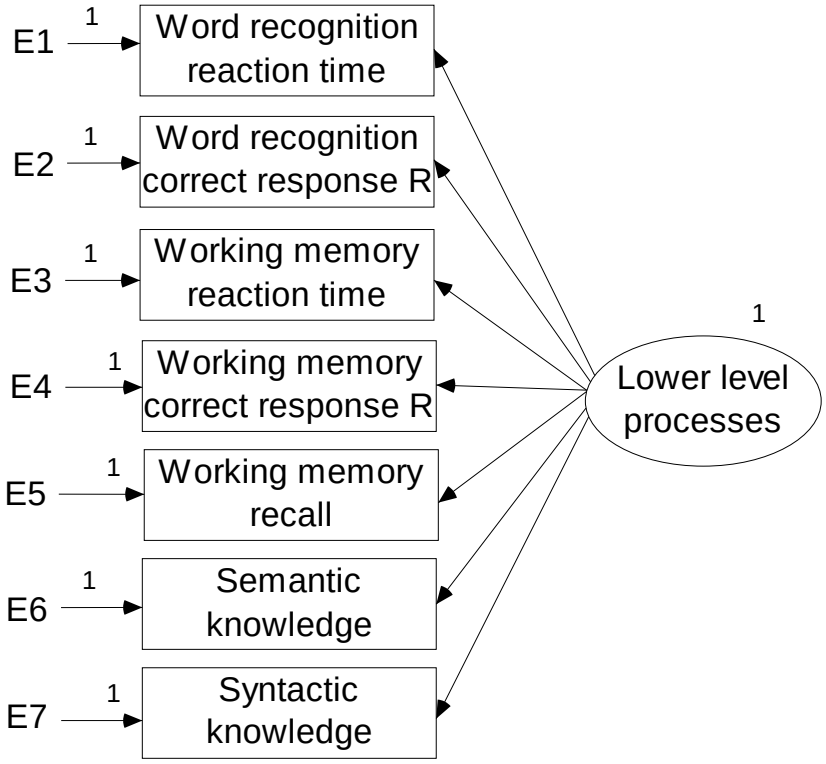


Figure 4.7 CFA model for the lower level processes

To examine whether the higher order CFA model 1 fit the data, a measurement model analysis of the lower-level processes was conducted first. As graphed in Figure 4.7, lower-level processes were indicated by seven observed variables, and thus there were 7*(7+1) = 28 unique pieces of information. The measurement errors of the seven observed variables were scaled to 1.0, and the variance of the factor, lower-level processes, was standardized and fixed to 1.0. The number of parameters requiring estimation was 14, i.e., seven variances of the nine measurement errors and seven factor

loadings. Therefore, the model was over-identified because the number of unique pieces of information was greater than the number of parameters requiring estimation. The degrees of freedom were (28-14) = 14. The same correlation matrix and scaled standard deviation used for the nine-observed-variable one-factor CFA model were employed. The analysis in Mplus 6.1 converged an admissible solution. Values of selected fit indices were χ^2 (14) = 58.105, $p < .05$, CFI = .677, SRMR = .094, and RMSEA = .139 with the 90% confidence interval .103 – .176. The only favorable index was SRMR, which was lower than the cut-off value of .10. Overall, the results indicated poor fit of the model. The Lagrange Multiplier test was conducted to examine whether model fit could be improved.

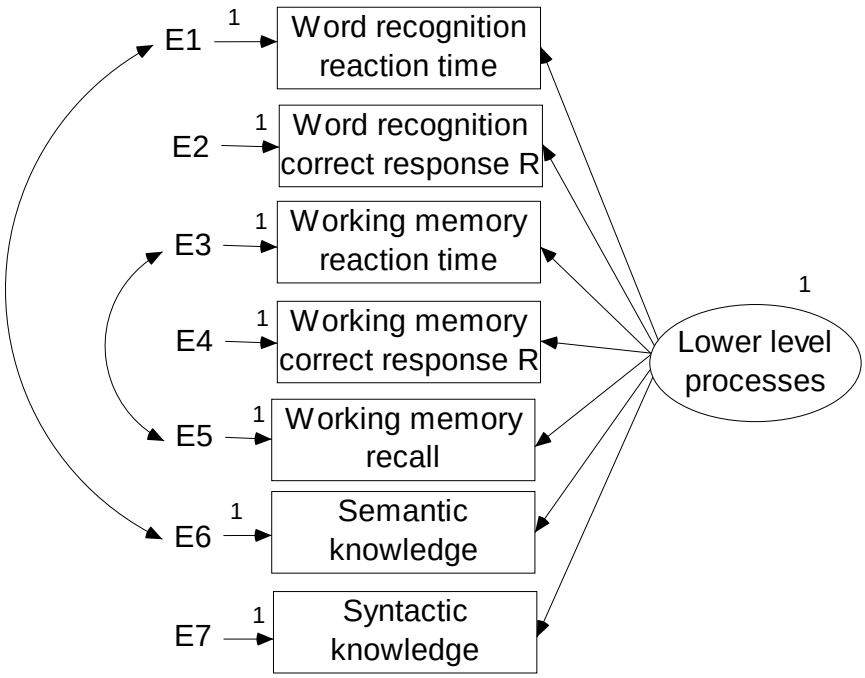


Figure 4.8 Respecified CFA model for lower-level processes

Six measurement error variance correlations were suggested. Two theoretical meaningful corrections, which would also lead to the largest decrease in chi-square values, were added to the model. Similar to the nine-observed-variable CFA model for reading ability, these correlations were the error variance correlation between word recognition reaction time and semantic knowledge, as well as that between working memory reaction time and working memory recall (see Figure 4.8). The degrees of freedom decreased from 14 to 12. The results revealed improvement of overall model fit. Values of selected fit indices were χ^2 (12) = 12.303, p = .421, CFI = .998, SRMR = .047, and RMSEA = .012 with the 90% confidence interval .000 – .018. These results indicated good fit of the respecified measurement model for lower-level processes.

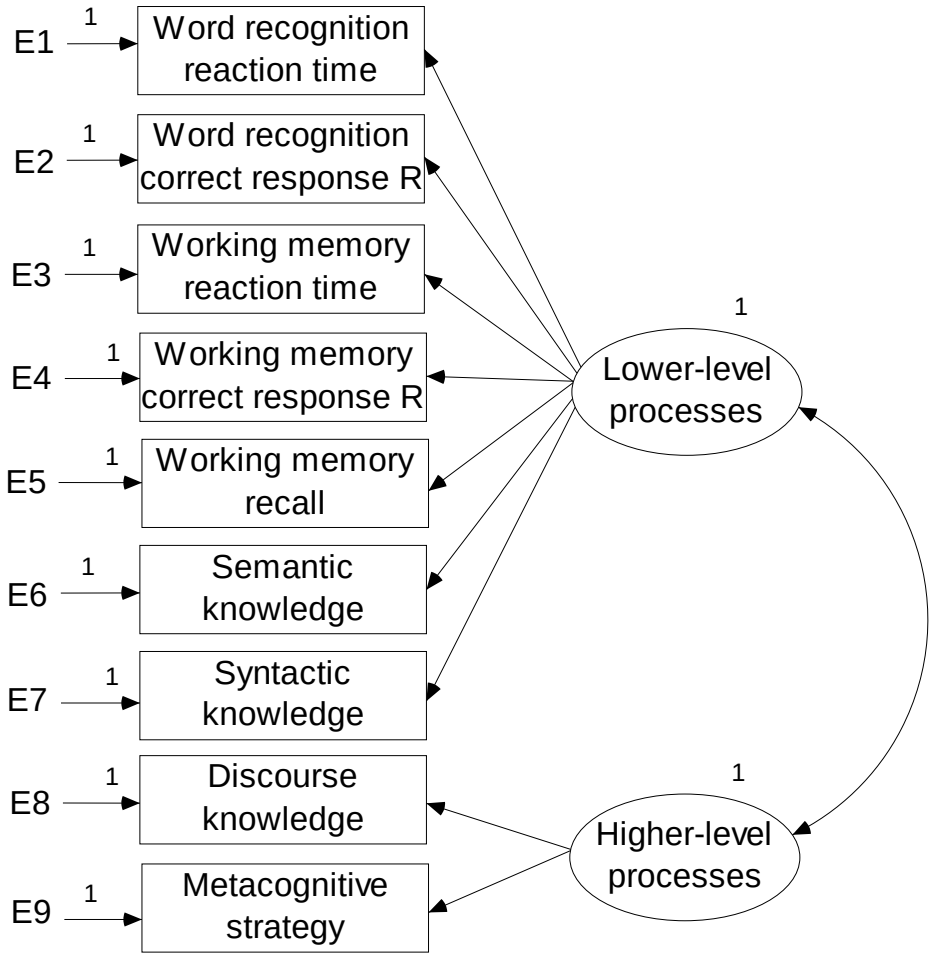


Figure 4.9 Two-factor (lower-level and higher-level processes) CFA model

The higher-level processes CFA model was locally unidentified because it had only two observed variables. Therefore, a separate confirmatory factor analysis was not conducted for the higher-level processes.

To test whether lower-level and higher-level processes were two dimensions of reading ability, the intercorrelation between them was examined. If they were highly

intercorrelated, it would not have been appropriate to view them as two separate dimensions of the factor of reading ability.

As shown in Figure 4.9 the two factors, lower-and higher-level processes, were allowed to covary. There were 45 unique pieces of information (9*[9+1]/2 = 45). The measurement errors were scaled to 1.0, and the variance of the factors, lower-level processes and higher-level processes, were standardized and fixed to 1.0. The number of parameters requiring estimation was 21, namely, nine variances of the nine measurement errors and nine factor loadings, two measurement error variance correlations, one factor correlation. Therefore, the model was over-identified, and the degrees of freedom were 45-21 = 24. The same correlation matrix and standard deviations as used in the nine-observed-variable CFA model for reading ability was utilized. The results of the analysis in Mplus showed that the latent variable covariance matrix was not positive definite. The latent variable correlation was 1.06, which was greater than 1.00. The problem involved the latent variable of higher-level processes, probably because it had only two observed variables.

In sum, the higher order CFA model 1 was not appropriate to serve as the baseline measurement model of reading ability for the purpose of structural model analysis. At this point, the nine-observed-variable one factor model was the best model. The following section intends to test whether another higher order model, which is based on the conceptualization of reading ability by Weir (2005), would be a competing, favorable measurement model of reading ability.

4.7.3 Higher order CFA model 2 for reading ability

As shown in Figure 4.10, the higher order CFA model 2 for reading ability was indicated by two factors: executive processes and executive resources. Word recognition reaction time and correct response rate, working memory reaction time and correct response rate, recall, and metacognitive strategy were indicators of executive processes, whereas semantic, syntactic, and discourse knowledge served as indicators of executive resources.

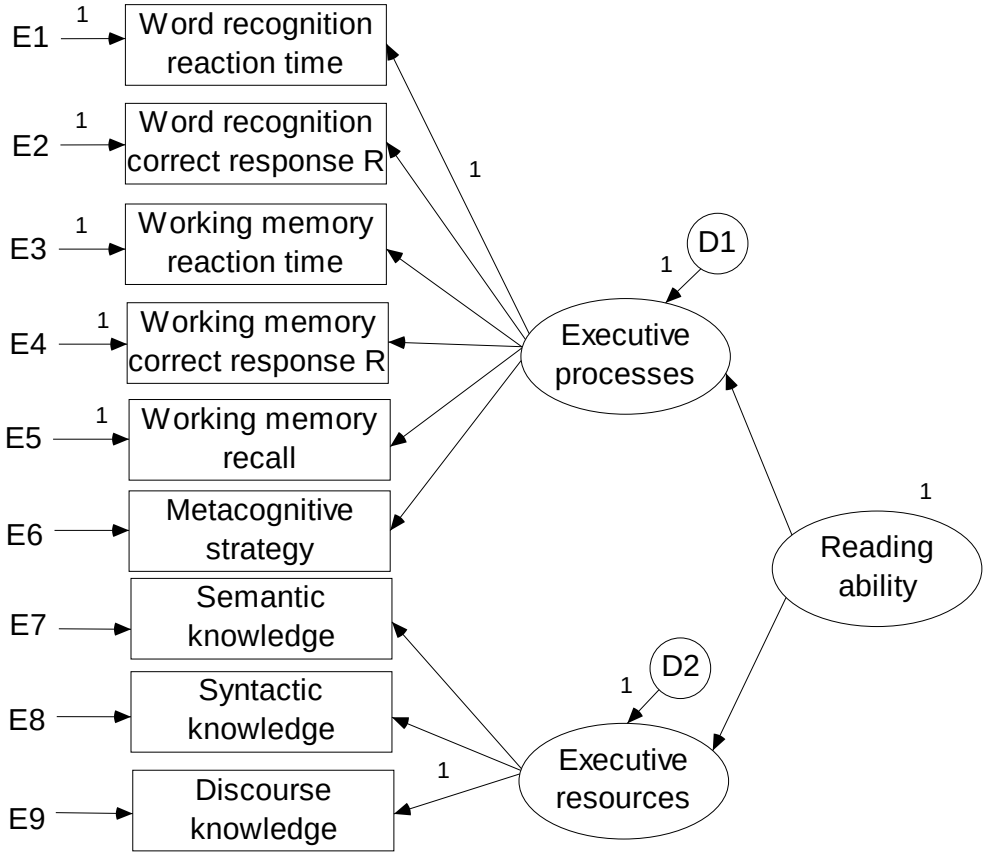


Figure 4.10 Higher order CFA model 2

A measurement model analysis of the executive processes was conducted first. As shown in Figure 4.11, executive processes had six observed variables, namely, word recognition reaction time, word recognition correct response rate, working memory reaction time, working memory correct response rate, recall, and metacogntive strategy. There were 6*(6+1) = 21 unique pieces of information. The measurement errors of the six observed variables were scaled to 1.0, and the variance of the factor, executive processes, was standardized and fixed to 1.0. The number of parameters requiring estimation was 12, i.e., six variances of the six measurement errors and six factor loadings. Therefore, the model was over-identified and the degrees of freedom were (21-12) = 9.

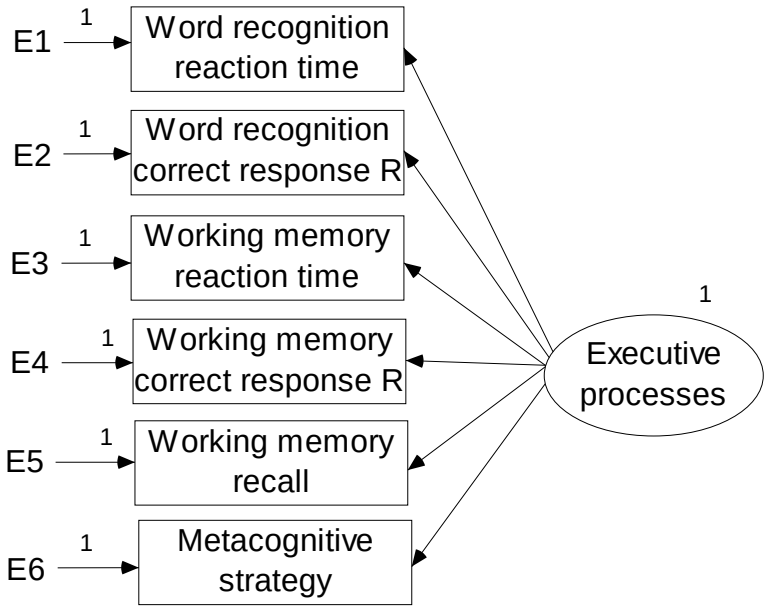


Figure 4.11 CFA model for executive processes

The same correlation matrix and standard deviations for the nine-observed-variable one-factor CFA model for reading ability was used for the executive processes CFA model analysis. The analysis in Mplus 6.1 converged an admissible solution. Values of selected fit indices were χ^2 (9) = 7.353, p = .600, CFI = 1.00, SRMR = .038, and RMSEA = .000 with the 90% confidence interval .000 – .076. The results indicated good fit of the model.

The latent variable, i.e., executive resources, had three indicators, namely, semantic knowledge, syntactic knowledge, and discourse knowledge. The measurement model for executive resources was just identified.

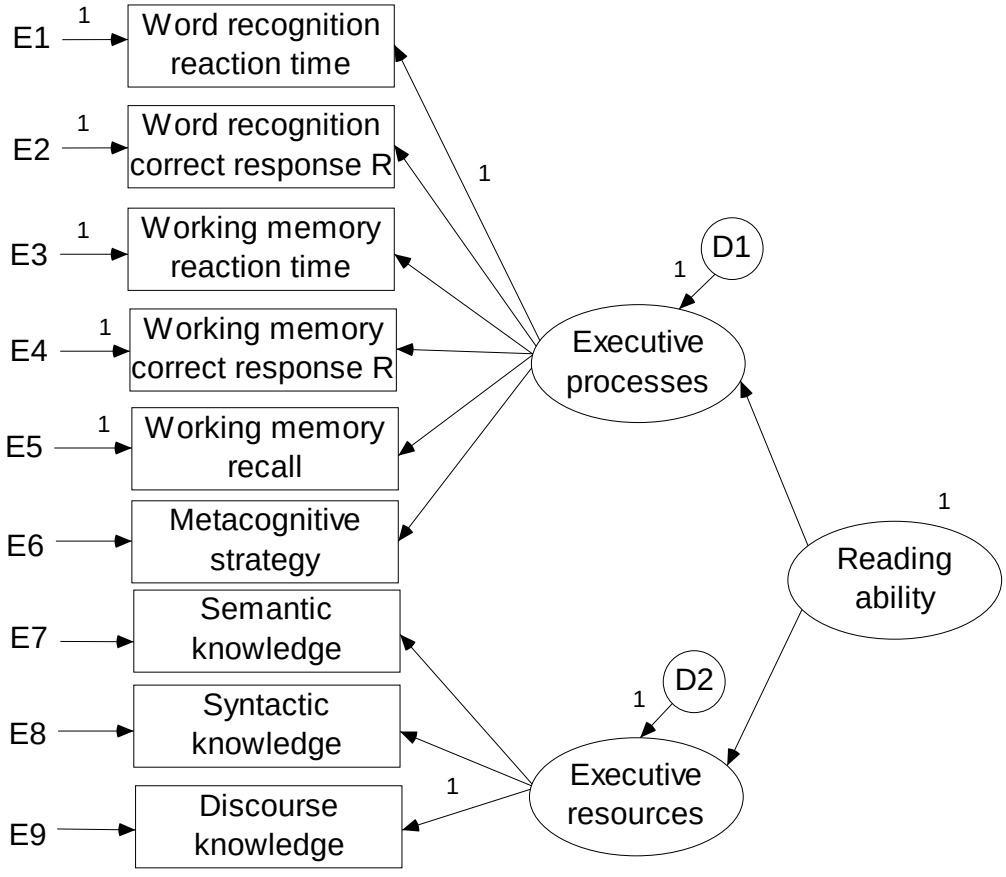


Figure 4.12 Two-factor (executive processes and resources) CFA model

The intercorrelation between the two latent variables, executive processes and executive resources, was tested. As shown in Figure 4.12 the two factors were allowed to covary. There were 45 unique pieces of information 9*[9+1]/2 = 45. The measurement errors were scaled to 1.0, and the variance of the factors, executive processes and executive resources, were standardized and fixed to 1.0. The number of parameters requiring estimation was 19, namely, nine variances of the nine measurement errors, nine factor loadings, and one factor correlation. Therefore, the model was over-identified, and the degrees of freedom were 45-19 = 26. The same correlation matrix and standard deviations as used in the nine-observed-variable CFA model for reading ability were utilized. The analysis in Mplus 6.1 converged an admissible solution. Values of selected fit indices were χ^2 (26) = 51.378, p = .002, CFI = .890, SRMR = .069, and RMSEA = .077 with the 90% confidence interval .045 – .108. The chi-square test was significant at the .001 level, which implied that the null hypothesis, i.e., the model fits the data, could not be retained. The value of CFI was smaller than the cut-offvalue of .96, which implied poor fit. SRMR was smaller than the normally used criterion value of .10, which implied that a good fit was associated with the model. RMSEA was .077, which was larger than the pre-determined value of .06, suggesting poor fit. Overall, this higher order CFA model was not as good a fit as the nine-observed-variable one-factor CFA model for reading ability.

Interestingly, the results showed that the correlation between executive processes and executive resources was .570, which implied that these two latent variables were good enough to serve as two dimensions of reading ability. Compared with higher order CFA model 1, which categorized the nine-observed-variables into lower-level and higher-level processes, higher order CFA model 2 captured the dimensions of reading ability better, judging by the values of the model fit indices and the correlation between the two latent variables.

In conclusion, the three competing CFA models for reading ability were not as good in terms of fit indices as the nine-observed-variable one-factor CFA model. Regarding the six-observed-variable CFA model, which excluded word recognition correct response rate, working memory reaction time, and working memory correct response rate, the analysis results veiled the underlining effect of reading ability on word recognition and working memory. Although the three excluded observed variables were not generally used as indicators of word recognition or working memory, they were stronger indicators of word recognition and working memory than the two customarily utilized indicators, i.e., reaction time for word recognition and recall for working memory. The nine-observed-variable CFA model better captured the effect of reading ability on word recognition and working memory.

Higher order CFA model 1 was the worst among the three competing CFA models. It had the poorest model fit indices and the correlation between the two latent variables, lower-level processes and higher-level processes, was not meaningful. As

regards higher order CFA model 2, although the correlation between the two latent variables, executive processes and executive resources, was reasonable, the model indices suggested poor model fit. Therefore, the nine-observed-variable CFA model was favored as the baseline model for the structural model analysis.

4.8 FULL LATENT VARIABLE STRUCTURAL MODEL ANALYSIS

The core research question of the present study aims to explore the degree to which test-takers' performances on the CET reading section are underlined by their reading abilities. To fulfill this research goal, the nine-observed-variable one-factor model (see Figure 4.3) was used as the baseline model of the measurement model for reading ability. Test-takers' performance was indicated by only one variable, which was their scores on the CET reading section. Although the CET reading section comprises three parts, i.e., fast reading, reading in depth, and a cloze with a word bank, only the total score in the reading section was accessible to the researcher of this study. Thus the test takers' performance had only one indicator, which raised the problem of local under-identification. To deal with this problem, the measurement error variance of the scores in the CET reading section was constrained. This value was calculated by one minus reliability and multiplied by variance of the CET scores, since one minus reliability was the proportion of total variance in the indicator that was attributable to error. The total variance of the CET scores on the reading section was obtained by squaring its standard deviation. However, the reliability was not accessible to the researcher of the current study. As suggested by Kline (2005) an estimate of the proportion of error variance could

be based either on the researcher's experience with the measure or on results reported in the research literature. For this study, the reliability documented by the CET committee was employed for the estimate of the error variance of the single indicator of the scores in the CET reading section. According to Yang & Weir (1998), the reliability of the CET excluding the writing section was .90. As presented in Table 4.10, the mean of the CET scores was 203.99, and its standard deviation was 33.97. To narrow the differences among the variances of observed variables so as to facilitate the iterative estimation process, the standard deviation of the CET scores was scaled to 3.40, with a scaling factor of 1/10. The new variance of the scores in the CET reading section was 3.40*3.40 = 11.56. The error variance of the scores was (1-.90) = 1.16. The following section outlines the steps of structural model analysis, including model specification, model identification, data summary, model estimation, and evaluation.

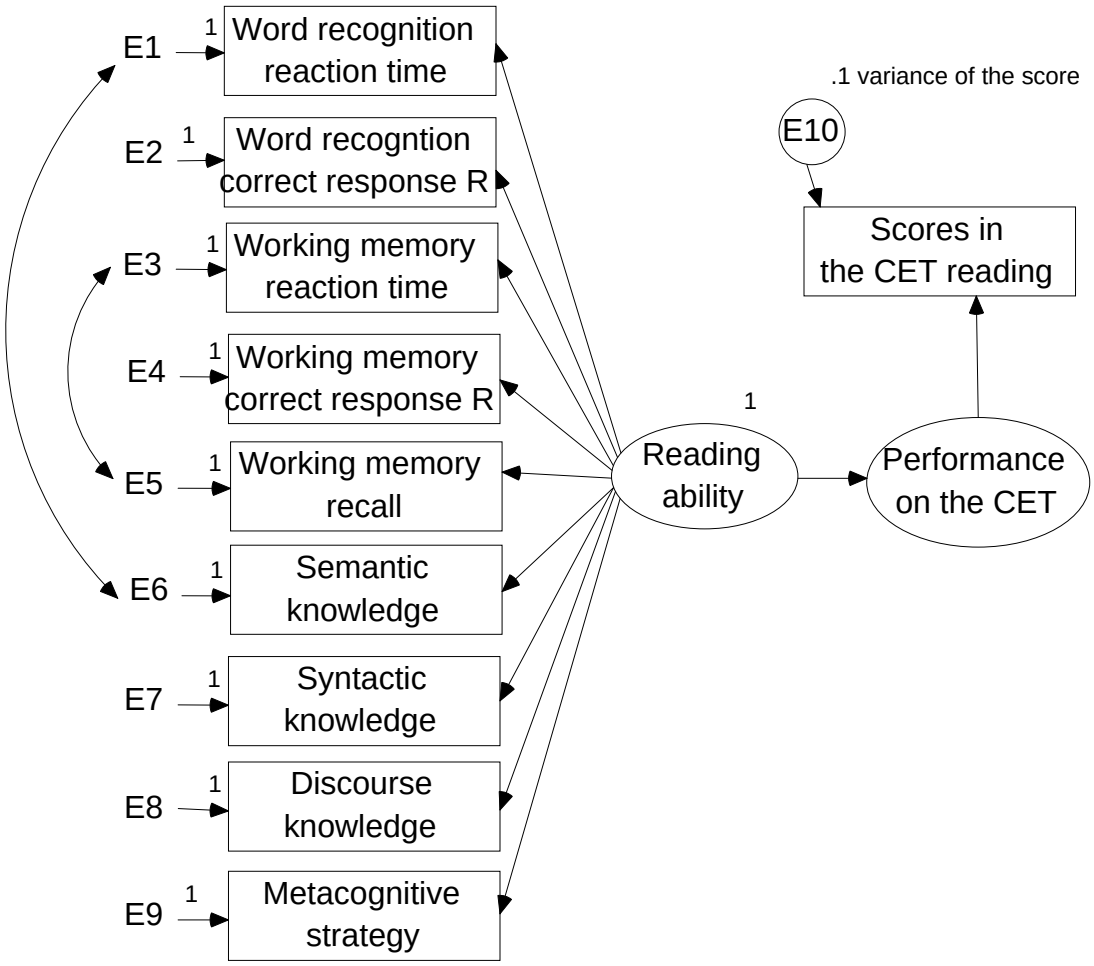


Figure 4.13 Structural model with the scores in the CET reading section

4.8.1 Model specification

As shown in Figure 4.13, the latent variable reading ability was indicated by nine observed variables, i.e., word recognition reaction times, word recognition correct response rates, working memory reaction times, working memory correct response rates, recall, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategies. Two paths of error variance correlations were added. The first

was the error variance correlation between word recognition reaction time and semantic knowledge. The second path was the error variance correlation between working memory reaction time and recall. The other latent variable, performance on the CET reading section, was indicated by the scores in the CET reading section. The latent variable of the performance on the CET was regressed on reading ability. The path value was used to examine the extent to which participants' performance on the CET reading section was underlined by their reading ability.

4.8.2 Model identification

There were 55 unique pieces of information in the standard deviations of the ten observed variables and their correlation matrix (10*[10+1]/2 = 55). An additional piece of information was that the error variance of the scores in the reading section was fixed at .1 of the variance of the scores. The measurement errors of the 10 observed variables were scaled to 1.0, and the variance of the factor, reading ability, was standardized and fixed to 1.0. The number of parameters requiring estimation was 23, i.e., nine error variances of the nine observed variables of reading ability, nine factor loadings between reading ability and its observed variables, two error variance correlations, one factor loading between performance on the CET reading section and the scores, the disturbance of the second latent variable, and the path between the two latent variables. Therefore, the model was over-identified because the number of unique pieces of information was larger than the number of parameters requiring estimation. The degrees of freedom were 56-23 = 33.

4.8.3 Data summary

A correlation matrix and the standard deviations of the 10 observed variables were used for analysis. Similar to the measure adopted in the measurement model, the standard deviations were scaled. Table 4.20 summarizes the scaled standard deviations and the correlations among the 10 observed variables.

4.8.4 Model estimation and evaluation

Mplus 6.1, the same software used for the measurement model analysis, was utilized for the structural model analysis. Table 4.21 lists the multiple fit indices. Values of selected fit indices were Chi-square χ^2 (33) = 50.532, p = .026, CFI = .946, SRMR = .058, and RMSEA = .057 with the 90% confidence interval .020 – .087. The chi-square test was significant at the .05 level, but it was not significant at the .01 level. This result implied that the model fit the data fairly well. The value of CFI was extremely close to the cut-off value of .95, which implied that the model had good fit. SRMR was smaller than the normally used criterion value of .10, which also implied that a good fit was associated with the model. RMSEA was smaller than the pre-determined value of .06, suggesting good model fit. Furthermore, the lower bound of the 90% confidence interval of RMSEA was smaller than .05, so the null hypothesis of close approximate fit could be retained. Overall, the results indicated that the model fit the data well.

Table 4.20 Correction Matrix and standard deviation for the ten observed variables

	WRRT	WRCR	WMRT	WMCR	WMRC	SEMK	SYNK	DISK	METAS	CET
Scaled SD	2.63	9.27	6.40	8.77	9.3	3.12	3.52	5.36	12.05	3.4
WRRT	1									
WRCR	-.204*	1								
WMRT	.228**	-.146	1							
WMCR	-.223*	.130	-.101	1						
WMRC	-.201*	.189*	-.394**	.165*	1					
SEMK	-.038	.268**	-.162*	.249**	.173*	1				
SYNK	-.198*	.280**	-.097	.242**	.070	.514**	1			
DISK	-.200*	.326**	-.164*	.271**	.063	.458**	.588**	1		
METAS	-.136	.151	-.163*	.115	.167*	.248**	.144	.200*	1	
CET	-.116	.313**	-.279*	.229**	.323**	.471**	.517**	.536**	.270**	1
Reliability		.63		.50		.86	.74	.80	.83	

Notes: * correlation is significant at the 0.05 level (2-tailed); ** correlation is significant at the 0.01 level (2-tailed); WRRT for word recognition reaction time; WRCR for word recognition correct response rate; WMRT for working memory reaction time; WMCR for working memory correct response rate; WMRC for working memory recall; SEMK for semantic knowledge; SYNK for syntactic knowledge; DISK for discourse knowledge; METAS for metacognitive strategy use; the reliability of semantic knowledge is the average of the reliabilities of the authentic words in the three Yes/No tests.

Table 4.21 Model fit indices for the structural model

Index	χ^2	*df*	*p*	CFI	RMSEA	90C.I. of RMSEA	SRMR
Value	50.53	33	.026	.946	.057	.020 - .087	.058

4.8.5 Parameter estimates interpretation

The path between the two latent variables, reading ability and test performance on the CET reading section, was reported first because this value reflected the degree to which the test performance was underlined by actual reading ability. As shown in Figure 4.14, the path estimate between reading ability and performance on the CET reading section was .75, which meant that each standard deviation increase in reading ability would lead to a .75 standard deviation increase in test performance of reading. The squared regression coefficient (R^2) was .565, which indicated that 56.5% of the variance of test performance was explained by the directeffect of reading ability. Both the path value and the R^2 were significant at the .001 level. These results implied that participants' performance on the CET reading section was to a large degree underlined by their reading ability. Furthermore, the factor loading from test performance to the scores in the CET reading section was .95, which meant that the scores were a good indicator of test performance. In conclusion, it is largely justified that we use participants' scores in the CET reading section to draw inferences about their reading ability. Participants who obtained higher scores in the CET reading section indicated that they had higher reading ability than those who scored lower.

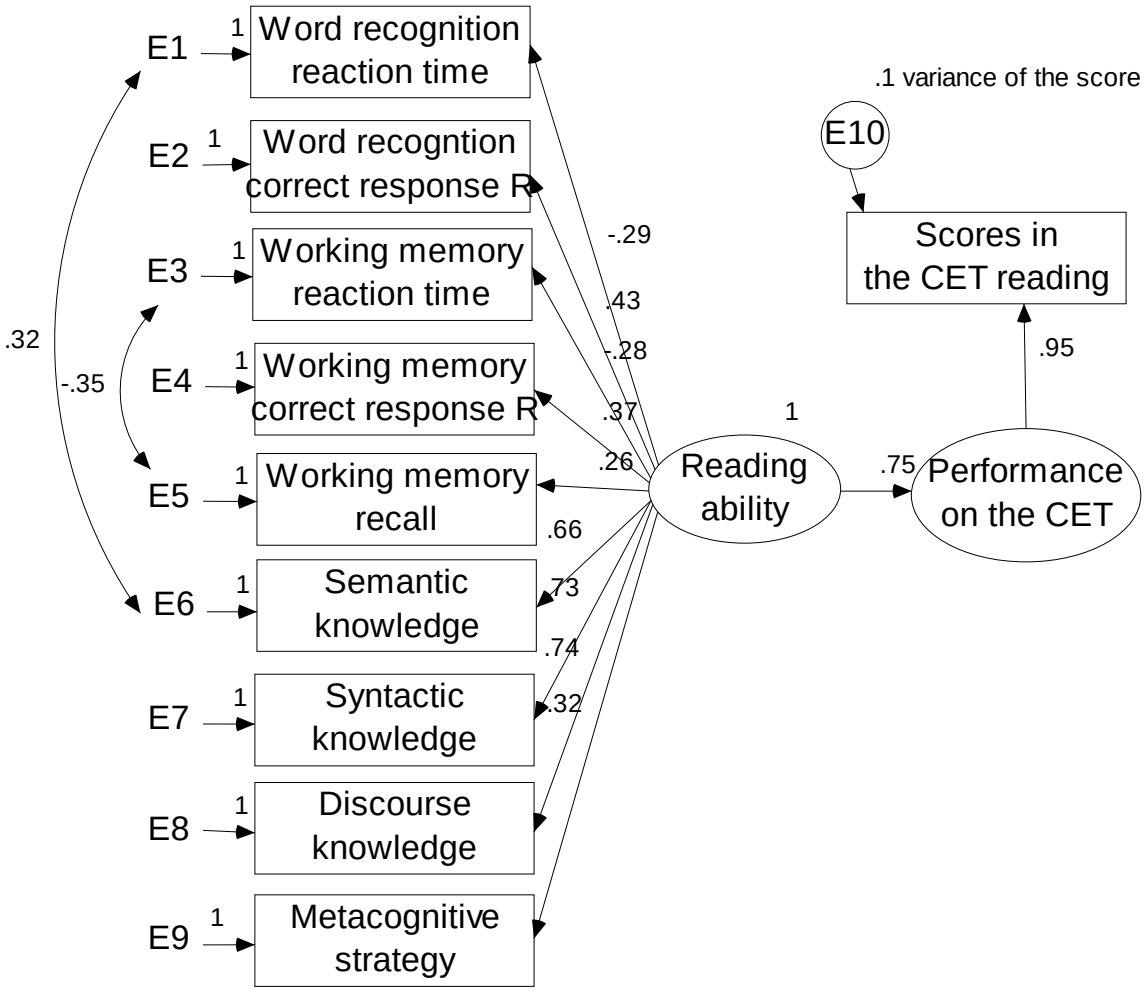


Figure 4.14 Structural model with standardized estimates

The parameter estimates between reading ability and their observed variables were slightly different from the values in the respecified measurement model of reading ability. As shown in Figure 4.14, all of the factor loadings were positive except those of word recognition reaction time and working memory sentence judgment reaction time. As discussed in the measurement model analysis, these two negative factor loadings were reasonable. First, readers with higher reading abilities tend to be faster in word

recognition reaction, thus their relationship was negative. For the same reason, readers with higher reading abilities tend to be more efficient in sentence processing which result in shorter reaction times.

Table 4.22 R^2 estimates of the structural model

Variable	R^2 estimate	*P* value
Word recognition reaction time	.083	.076
Word recognition correct response rate	.188	.002
Working memory reaction time	.076	.083
Working memory correct response rate	.134	.015
Working memory recall	.067	.108
Semantic knowledge	.438	.000
Syntactic knowledge	.535	.000
Discourse knowledge	.543	.000
Metacognitive strategy	.100	.042
CET scores	.899	.000
Latent variable reading performance	.565	.000

Similar to the pattern of the measurement model of reading ability, three factor loadings stood out among the nine observed variables, namely, the direct effect of reading ability on semantic knowledge, syntactic knowledge, and discourse knowledge. As shown in Figure 4.14, the estimates were .67, .73, and .74, respectively. As shown in

Table 4.22, the squared regression coefficient for semantic knowledge was .438, which indicated that 43.8% of the variance in semantic knowledge could be explained by reading ability. The values for syntactic and discourse knowledge were 53.5% and 54.3%, respectively. More than half of their variance could be accounted for by the latent variable of reading ability. These results implied that these three variables were still very good indicators of reading ability in the structural model.

The path from reading ability to word recognition correct response rate was .433, which implied that each standard deviation increase in reading ability would lead to a .433 standard deviation increase in word recognition correct response rate. Comparatively, the path from reading ability to word recognition reaction time was smaller, -.29, although significant at the .01 level. This result implied that word recognition correct response rate was a better indicator than word recognition reaction time, although reaction time is customarily used as the sole indicator of word recognition. The paths from reading ability to working memory reaction time, correct response rate, and recall were -.28, .37, and .26, respectively. The highest was the factor loading of working memory correct response rate, although recall is generally used as the only indicator of working memory capacity. These three factor loadings were significant at the .01 level. The squared regression coefficients of these three observed variables were .076, .134, and .067, respectively. Similar to the estimates of word recognition, the highest value was the R^2 of working memory correct response rate, rather than reaction times or recall. Reading ability accounted for 13.4% of the variance in working memory correct

response rate, which was significant at the .05 level. However, reading ability explained less than 10% of the variance of working memory reaction time and recall. Neither of them was significant at the .05 level.

The factor loading of metacognitive strategy was .32 and its R^2 was .100, both significant at the .05 level, although these values were very small compared with the estimates related to semantic, syntactic, and discourse knowledge. These results implied that metacognitive strategy was a fair indictor of reading ability.

The parameter estimates of the error variance correlations were similar to the measurement model of reading ability. The error variance correlation between working memory reaction time and recall was -.35 ($p < .01$), which implied that shorter reaction times in sentence processing resulted in better recall of the words attached to the sentences. The negative correlation between these two paths represented the competition for the limited resources of working memory between processing and memorizing.

The error variance correlation between word recognition reaction time and semantic knowledge was .32 ($p < .01$), which implied that their measurement errors might come from a shared source. Considering that both the measurement of word recognition and semantic knowledge involved the use of pseudowords, the measurement errors might arise in part from the use of this measurement approach.

In conclusion, this chapter first examined the descriptive statistics of the data from the six measurements, namely, word recognition, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy. The

means and standard deviations were summarized. The univariate normality of the data, which is the assumption of the maximum likelihood estimation, was also scrutinized. The results showed that the data of each measurement were normally distributed and therefore appropriate for the structural equation modeling analysis.

Second, based on participants' responses to individual items, the internal reliabilities of the six measurements were presented. The reliabilities of semantic, discourse task A and metacognitive strategy measurements were above .80, which indicate that the data yielded from these measurements were consistent across items. The reliability of syntactic knowledge measurement was .74, indicating that the instrument was reliable enough to generate data for the measurement of syntactic knowledge. The internal reliability of word recognition was .63. The two lowest reliabilities were those of the working memory measurement and the discourse knowledge measurement Task B, .50 and .42, respectively. These two lower reliabilities were most likely due to the limited number of items. The working memory task comprised only 28 items, which was further divided into two parallel sets. The discourse knowledge measurement Task B only had six mini passages. Given the large total number of items, 323 for each participant, an increase in the number of items in these two tasks would result in a change of research design. A possibility would be to split the whole test into two parts and to administer it in two different days.

Third, the CET scores in the reading section were reported, including the mean and the standard deviation. The normality of the data distribution was checked, and the

result showed that the scores were normally distributed and thus met the assumption of maximum likelihood estimation.

Fourth, the nine-observed-variable one-factor confirmatory model for reading ability was analyzed and respecified. The results indicated good model fit. Three competing confirmatory models for reading ability were analyzed and compared with the nine-observed-variable one-factor confirmatory model. The results implied that the nine-observed-variable one-factor confirmatory model was superior to the three competing models. Compared with the six-observed-variable CFA model, the nine-observed-variable one-factor confirmatory model illustrated that the effect of reading ability on word recognition and working memory was mainly on the indicators of their correct response rates. However, when correct response rates were excluded from the model, the directeffect of reading ability on word recognition and working memory was concealed. The nine-observed-variable CFA model was superior to the two higher order CFA models when comparing their model fit indices.

Finally, the structural model was analyzed, which included the respecified nine-observed-variable one-factor measurement model of reading ability and the scores in the CET reading section. The results indicated that the structural model showed good model fit indices. The path from reading ability to test performance in the CET reading section was .75, which implied that participants' test performance and scores in the CET reading section were strongly underlined by their actual reading ability. Moreover, 56.5% of the variance of test performance could be explained by reading ability. Given that the CET

reading section accounted for only 35% of the whole CET, participants' reading ability variance could not be fully captured. The test time and number of items assigned to the reading section was limited. Therefore, the path value from reading ability to test performance revealed in the present study was strong enough to provide positive evidence for the construct validity of the CET reading section. The scores in the CET reading section were justifiable to a large degree for use in drawing inferences about participants' reading ability.

Chapter 5 Conclusion

Scores on educational measurements are generally meaningless without thoughtful interpretation. The interpretations of a particular score or an observation are usually based on a number of assumptions that are implicit in validity arguments. In L2 assessment, the implicit assumptions may include "the items of a test are a good sample of tasks", "the test has high reliability", "raters' scoring is consistent", "the scores are attributed to test-takers' language ability", "test-takers who obtained high scores on Test A tend to score high on Test B which measures the same construct as Test A". The plausibility of a validity argument will be strengthened if these assumptions are substantiated by evidence.

The present study focused on the plausibility of the argument that the scores of the CET reading section could be used as a measurement of test-takers' reading ability. In other words, this study aimed to examine the construct validity of the CET reading section; specifically, the extent to which scores on the CET reading section could be used as a measurement of test-takers' English reading ability.

Employing an interpretative argument approach (Toulmin, 1958, 2003; Kane, 1992, 2001; Mislevy et al., 2002, 2003; Chapelle et al., 2008) and guided by Weir's (2005) framework for sources of evidence that support or cast doubt on the interpretation of reading tests, the present study focused on the assumption that test-takers' performances on the CET reading section are attributed to their reading ability. The following section will summarize the research results of this primary question as well as the study's findings with respect to the process of modeling reading ability.

5.1 CONSTRUCT VALIDITY OF THE CET READING SECTION AS REVEALED IN THE PRESENT STUDY

The two interrelated research questions are:

1. To what extent do test-takers' reading abilities account for their performances on the reading section of the CET?
2. To what extent can the variance of test-takers' scores on the reading section of the CET be explained by their reading ability?

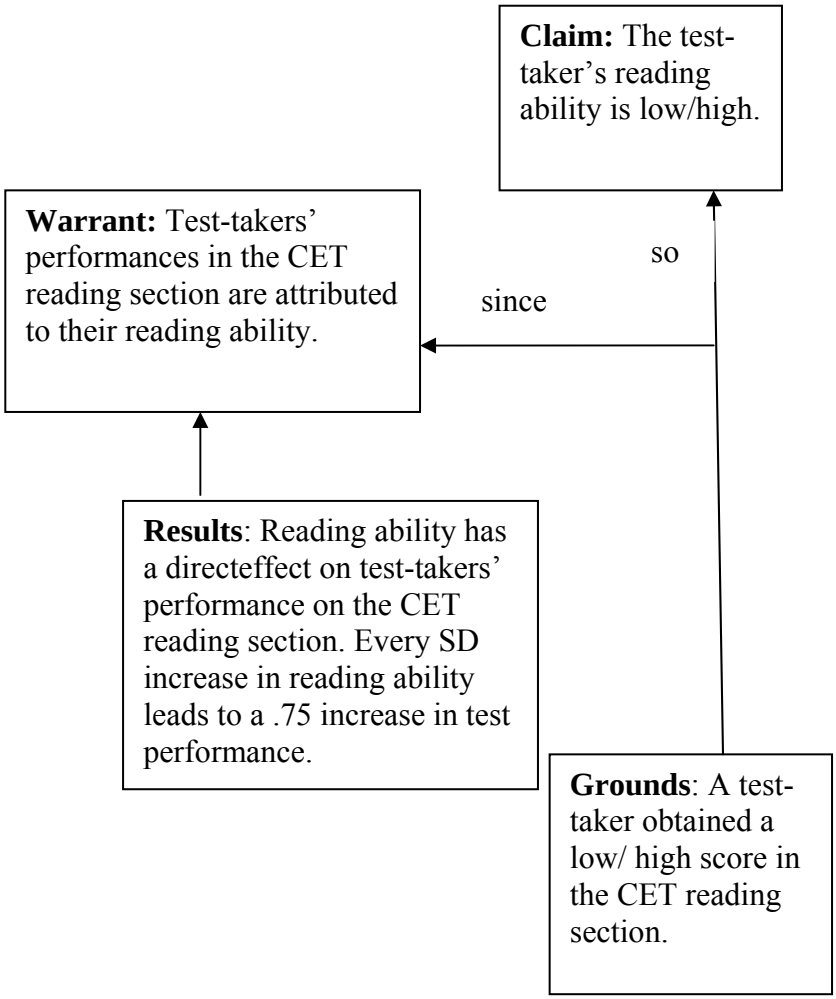


Figure 5.1 The interpretative argument and the research results

Utilizing structural equation modeling, the results of the present study found that reading ability has a directeffect on test-takers' performance on the CET reading section. Every SD increase in reading ability leads to a .75 increase in test performance (see Figure 4.14), and reading ability accounts for 56.5% of the variance of test performance on the CET reading section. Since the CET is an English language proficiency test, in which listening, reading, writing, and translating component skills are incorporated, reading comprises only 35% of the total score and 35% of the total test time. Therefore, it could be concluded that test-takers' reading abilities account to a large degree for their performances on the reading section of the CET, and a large amount of variance of test-takers' scores on the reading section of the CET could be explained by their reading ability. The scores on the CET reading section are largely justifiable for use in drawing inferences about test-takers' reading ability, which was operationalized by reading processes, linguistic knowledge, and the metacognitive strategy of reading.

The positive evidence for the construct validity of the CET reading section reported in the present study is in accordance with the findings from Yang & Weir (1998) and Jin & Wu (1998). This positive evidence in favor of the reading section of the CET could serve to strengthen teachers, students, and other CET stakeholders' belief in the use of the CET scores as a measure of reading ability and overall EFL proficiency although the validity of the entire CET was not examined in this study. Teachers would be more confident in placing emphasis on the development of students' reading ability and

language proficiency in the classroom even under the pressure of accountability of college foreign language education.

5.2 IMPLICATIONS FOR VALIDATION STUDIES OF THE CET

Exploring reasons for the paucity of the studies on the CET is not a goal of the present research. However, the lack of necessary information, either in the form of raw scores or summary data, renders the exploration of validity evidence impossible in most cases. Consequently, the evaluation of the CET has become the responsibility, or the privilege, of the CET committee.

By examining the degree to which test-takers' variance in the scores on the reading section is underlined by their variance in the reading section, the present study sheds light on the research designs for teachers and researchers who have no access to the CET raw scores. He & Dai's 2006 study on the construct validity of the CET Spoken English Test also focused on the examination of test-takers' speaking performance without access to participants' raw scores. Employing similar research methods, teachers could conduct studies on the theory-based validity of the listening, translation, and writing sections, or even on the test as a whole. Broadly speaking, teachers would be able to design studies on the concurrent validity of the CET individual sections or on the entire test with only the total scores and section scores, which are accessible to test-takers' instructors. In this way, teachers could play a more active role in the exploration of the validity evidence for the CET, which would contribute to the development of the CET and language assessment research in China.

Estimating reading ability constitutes a critical step in the process of exploring answers to the above-listed research questions. The following section will summarize some major findings in the confirmatory analysis of reading ability.

5.3 THE COMPONENTS OF READING ABILITY

The results of the current study have shown that reading ability could be indicated by word recognition, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy. These six components of reading yield nine observed variables for the latent variable of reading in the confirmatory factor analysis, because word recognition has two indices, i.e., reaction time and correct response rate, while working memory has three indices, i.e., reaction time, correct response rate, and recall. This baseline confirmatory model for reading ability is called the nine-observed-variable one-factor model.

Three other confirmatory models were analyzed and compared with the nine-observed-variable one-factor model. The first is a six-observed-variable model, in which word recognition correct response rate, working memory reaction time and correct response rate have been excluded. The second and third models are both higher order confirmatory models, in which the latent variable of reading ability is indicated by two other latent variables. In the second model, the first-order, two latent variables are lower-level processes and higher-level processes, whereas in the third model, the first-order two latent variables are executive processes and executive resources. The nine-observed-variable one-factor model is superior to the two higher order confirmatory models

judging by the model fit indices. Although the six-observed-variable model exhibits the same good model fit indices as the nine-observed-variable one-factor model, the latter has been chosen as the baseline model for the structural equation model because the nine-observed-variable one-factor model represents more clearly the relationship between reading ability and its components.

The three strongest indicators are semantic, syntactic, and discourse knowledge. Their factor loadings are .67, .73, and .74, respectively. The squared regression coefficients of these indicators are also the three highest among the nine observed variables, i.e., 43.8%, 53.5%, and 54.3%, respectively. Each of these estimates is significant at the .001 level. By contrast, the factor loadings for word recognition, working memory, and metacognitive strategy are much smaller, ranging from .288 to .433, although these factor loadings are significant at the .001 level. Furthermore, the squared regression coefficients for word recognition reaction time, working memory reaction time, and recall are not significant even at the .05 level. Word recognition correct response rates, working memory correct response rates, and metacognitive strategy are significant at the .05 level. Word recognition correct response rates and working memory correct response rates are stronger indictors than their reaction times.

The contrast between the significant roles of linguistic knowledge and reading ability as well as the statistically marginal significance of word recognition and working memory has also been demonstrated in other studies that utilize the component skills approach (e.g., Haynes & Carr, 1990; Nassaji & Geva, 1999; Nassaji, 2003; Shiotsu,

2003; van Gelderen et al., 2004; van Gelderen et al., 2007). In Nassaji & Geva's study, the syntactic and semantic measures accounted for a substantial proportion of the variance of reading abilities regardless of their entry sequence, either first or last entering the regression equations. However, other components were very sensitive to the order of entry. In van Gelderen et al.'s 2004 study, even though vocabulary and grammar knowledge explained a substantial amount of the variance of reading comprehension, they found no unique contributions of processing speed components when linguistic and metacognitive knowledge had been controlled for.

In summary, L2 reading ability for the participants of this study is indicated by three clusters of components in terms of statistical estimates. The first is the robust linguist knowledge cluster, which includes semantic, syntactic, and discourse knowledge. The second is the power component of word recognition and working memory, namely, correct response rates of word recognition and working memory. The last cluster includes the weak and elusive indicators of speed components of word recognition and working memory, as well as metacognitive strategy.

5.4 IMPLICATIONS FOR L2 READING THEORIES

This section will focus on the implications of the findings for the component skills of reading, as well as issues related to L2 readers' reading automaticity.

5.4.1 The component skills of L2 reading

The three clusters of components are categorized according to their statistical values, in particular, factor loadings and squared regression coefficients. However, a closer examination of the power component of word recognition and working memory, which refers to word recognition correct response rates and working memory correct response rates, may reveal that this component could be categorized as linguistic knowledge. First, word recognition correct response rates involve the accuracy with which participants can translate visual letter strings into meaning. This skill is a major dimension of semantic knowledge. Second, working memory correct response rates entail the accuracy with which participants can make meaning of sentences, which is a natural result of syntactic knowledge. Syntactic knowledge is defined as how a reader knows about the way in which words and phrases are combined to form sentences in a language. Therefore, the ability of making meaning of sentences is a demonstration of syntactic knowledge.

Another issue is related to the categorization of metacognitive strategy use. Although the values of its factor loading and R^2 are similar to those of the speed component of word recognition and working memory, it is not appropriate to group them together. The speed component reflects how quickly and efficiently participants process words and sentences, whereas metacognitive strategy deals with the flexibility of a reader in selecting actions to achieve particular reading goals.

Based on the above analysis, it could be concluded that reading ability is largely indicated by linguistic knowledge, processing efficiency, and metacognitive strategy. This conceptualization of reading ability is similar to Weir's (2005) description of the construct of reading. According to Weir, executive resources and executive processes in addition to monitoring skill constitute the dimensions of reading ability. Weir's executive processes are loosely parallel to processing efficiency. However, in addition to the efficiency of word recognition and syntactic parsing, Weir's executive processes component has incorporated goal setting, which was categorized as metacognitive strategy in the present study. Weir's executive resources are parallel to the linguistic knowledge component of this study, but the former also includes pragmatic knowledge and sociolinguistic knowledge.

However, the argument that the three components of reading ability, namely, linguistic knowledge, processing efficiency, and metacognitive strategy, are based solely on statistical strengths and analysis. In future studies, exploratory factor analysis would be more appropriate for the exploration of the components of reading ability.

5.4.2 Lower-level processing efficiency of L2 reading

The results of the present study have shown that the factor loadings of word recognition reaction time and sentence processing reaction time in the working memory task are significant but very small, i.e., -.288 and -.276, respectively. The squared regression coefficients of these two indicators are small and not significant at the .05 level, which implies that variances of these indicators explained by reading ability are not

significant. Strictly speaking, reaction time is just one component of processing efficiency. The other component could be the quality of processing, which might be indicated by either automaticity or correct response rates. Segalowitz & Segalowitz (1993) employed the coefficient of variation of reaction time — the standard deviation of reaction time divided by mean reaction time — as an index of automaticity. However, for the present study, the three terms: processing efficiency, processing speed, and processing automaticity, are used interchangeably.

The weak relationship between reading ability and processing efficiency revealed in this study is in line with other studies that employ the component skills approach (e.g. Haynes & Carr, 1990; Nassaji & Geva, 1999; Nassaji, 2003; Shiotsu, 2003; van Gelderen et al. 2004; van Gelderen et al., 2007). This weak and statistically significant or nonsignificant relationship between reading ability and processing efficiency leads to diverse interpretations.

Haynes & Carr (1990) observed that different variables correlated highly with reading comprehension and reading speed measures. Speeded tests, such as visual matching of words, had higher correlations with the measure of reading speed than with reading comprehension. However, L1 reading comprehension and English writing system knowledge correlated highly with English reading comprehension but not with reading speed. Based on this observation, Haynes & Carr contended that reading comprehension and reading speed are not influenced by different variables.

Similarly, Shiotsu (2003), using exploratory factor analysis after he was unable to identify a reliable relationship between processing efficiency and L2 reading comprehension in the main study, found that the participants' performances were best explained by two latent factors: careful text processing power and semantic access efficiency.

Nassaji & Geva (1999) found a weak but statistically significant R^2 change in their multiple regression analysis. This result was interpreted to mean that lower-level processing efficiency made unique contributions to the measure of reading ability after the role of linguistic knowledge was accounted for. Nassaji & Geva concluded that the role of lower-level processes must not be neglected even in highly advanced L2 readers.

Van Gelderen et al. (2004) compared two models: a basic model and a model that fixed the regressions on the speed components to zero. Van Gelderen et al. found no significant change in model fit index. They interpreted the result as the speed components made no significant, unique contribution to the explanation of L1 and L2 reading comprehension. Supported by the large contribution of L1 to L2, they concluded that the L1 to L2 transfer theory of reading explains the participants' performances better than theories that stress the importance of L2 linguistic knowledge (Alderson, 1984; Clarke, 1979) or theories that emphasize the role of processing efficiency for successful L2 reading (Favreau & Segalowitz, 1983; Koda, 1996; Segalowitz, 2000).

This researcher would like to present three arguments regarding the weak relationship between the component of efficiency and reading ability, which was

indicated by the small but significant factor loadings and non-significant R^2 of word recognition reaction time and working memory reaction time.

The first argument is that processing efficiency may be more related to reading speed than to careful reading power, which is similar to the opinion of Haynes & Carr (1990) and Shiotsu (2003). This argument could have been strengthened if a closer relationship between processing efficiency and fast reading had been found. However, information about the participants' scores in fast reading was not accessible to the researcher of this study. The finding that word recognition reaction time has a higher correlation with working memory reaction time than with any other observed variables may imply that these two efficiency variables are underlined by a latent variable of reading speed.

The second argument is that the individual's differences in processing efficiency could be used as a variable to explain the differences in L2 reading ability, which is similar to Nassaji & Geva's (1999) contention. Although the factor loadings of word recognition reaction time and working memory reaction time are small, they are still significant at the .001 level.

The third argument involves a description of L2 reading ability in general. The weak relationship between processing efficiency and reading ability — revealed by the non-significant R^2 of word recognition reaction time and working memory reaction time in this study and by the finding that the statistical significance of lower-level processing efficiency is sensitive to the entry sequence of the variables in the multiple regression

analysis — might be a representation of the phenomenon that L2 reading is generally lacking automaticity and fluency. Although not as observable as in L2 speaking, the lower reading rates and lesser automaticity are a major difference compared with L1 reading. It is also one of the 12 differences between L1 and L2 reading that have been outlined by Grabe (2000). The inadequacy of automaticity and efficiency in L2 reading has been pointed out by other L2 reading scholars (e.g., Bernhardt, 1991; Geva et al., 1997; Haynes & Carr, 1990; Segalowitz et al. 1991).

It seems natural to extend the observation of the lack of automaticity in L2 reading to L2 reading instruction by suggesting remedial measures such as extensive reading and practice with word recognition skills. However, the researcher of this study will interpret the implications for L2 reading pedagogy somewhat differently.

5.5 IMPLICATIONS FOR L2 READING PEDAGOGY

Before interpreting the findings of this study with regard to implications for L2 reading instruction, two preliminary points need to be made at the outset. The first point is that L1 and L2 reading abilities have fundamentally different functions for readers. The second point is that L2 reading goals vary dramatically, which challenges Grabe's (2000) claim that the skilled L1 reader "is the end point of expertise that an L2 reader is aiming towards" (p. 227).

First, L1 reading ability influences the reader in almost every respect, schooling, daily life, work, communication with friends, and entertainment. However, the impact of L2 reading ability is much more restricted. For the English as foreign language learners

in China, L2 reading ability would most likely affect their educational prospects, future research, or business to an extent. L2 learners can make a normal living without L2 reading ability, but they could not do so without adequate L1 reading ability.

Second, the spectrum of L2 learning goals is larger than most foreign language teachers or researchers realize. For example, some learners of English in China just want to pass an exam so that they can get their high school diploma, some hope to achieve a higher score on the CET to have an edge in the job market, others plan to take the TOEFL and the GRE for the purpose of getting accepted to a good graduate school, and still others strive to enter academia or business in the United States or in other English speaking countries. For the latter group of L2 learners, it is appropriate to believe that their final reading goal should be to obtain the skill level of the L1 reader. For the overwhelming majority of English language learners, however, it is not advisable for teachers to assume that their L2 reading goal is to read at the native or L1 level.

Therefore, one implication of the findings of this study is that extensive reading should be treated as a compulsory course in order to develop a high L2 reading ability. For students in secondary foreign language schools, in which foreign languages are used as medium instruction languages, students in university foreign language departments, and foreign language teachers, it is essential to read extensively to develop L2 reading automaticity and fluency.

A second implication is that for the non-English major college students, i.e., the majority participants of the CET, English instruction is appropriate for focusing on

vocabulary, syntactic knowledge, and discourse knowledge. The three largest factor loadings of semantic, syntactic, and discourse knowledge can be translated as the reading ability of the participants are mostly indicated by their semantic, syntactic, and discourse knowledge. Therefore, knowledge in these three aspects is closely related to reading ability. However, it is not sensible to conclude that L2 reading instruction should emphasize the construction of these three knowledge bases.

Researchers tend to interpret findings from studies in two different ways: prescriptive and descriptive. For example, there is an observation that reading ability correlates more highly with syntactic knowledge than with any other variable, or that syntactic knowledge has the largest squared regression coefficient change in a multiple regression analysis, or that syntactic knowledge has the largest factor loading in a structural equation modeling analysis. Researchers who take a prescriptive perspective would interpret the finding as syntactic plays the most important role in reading ability. Researchers who take a descriptive perspective would conclude that there exists a closer relationship between reading ability and syntactic knowledge, but this does not imply that syntactic knowledge is the most important variable for the development of reading ability. It might be that the variances of other variables are too small to be detected, either because they are highly developed or not developed at all. The researcher of this study takes more of a descriptive perspective with respect to the research findings and resorts to L2 reading theories for the prescriptive need.

In conclusion, the implications of the present research for L2 reading pedagogy are twofold. If the students' L2 learning goal or requirement is to develop native like reading ability, it is essential to design an extensive reading curriculum that allow them to acquire reading automaticity and fluency. If the goal is to develop communicative reading ability, for instance to find scientific information, it is appropriate to focus L2 reading instruction on the build-up of linguistic knowledge. It is beneficial for L2 reading instructors to be aware of the importance of extensive reading and assign a reasonable number of reading tasks to students. However, it is not sensible to assume a native like L2 reading ability as the goal for every L2 learner. Developing a communicative foreign language reading ability is a far more rational objective than that of acquiring an L1 like reading ability. As Grabe (2000) states "This bottleneck for reading processing is not easily circumvented and may take many years to overcome, if it ever is overcome" (p. 245). Therefore, it is of the same importance for teachers to realize the effort and time involved in developing automaticity and fluency in L2 reading. This balanced information may help foreign language teachers and educators to design a more productive curriculum and make the best use of students' time and educational resources.

5.6 IMPLICATIONS FOR L2 READING ASSESSMENT

Grabe (2000) has pointed out that the findings of reading research have little impact on reading assessment. While reading research focuses more on the cognitive processes in reading comprehension, reading assessment usually attaches importance to the product of comprehension. The question of how to bridge the gap between these two

research fields deserves consideration from both reading researchers and language assessment researchers. This section is an attempt to narrow the gap between reading research and reading assessment.

The results of this study have revealed that discourse knowledge is the strongest indicator of reading ability. As shown in Figure 4.14, discourse knowledge has the greatest factor loading and the largest squared regression coefficient. However, measurement of discourse knowledge has not been incorporated into mainstream foreign language assessment, such as the CET or the TOEFL. The researcher of this study does not intend to convey the notion that every reading process should be measured individually but rather that the most important component should be emphasized in reading assessment.

Rational deletion cloze and identification of top-level organization are the two instruments used to tap discourse knowledge. These two types of items could be considered as candidates for new tasks in reading assessment.

Another implication of the findings of the present study involves the component of the processing efficiency in reading. Both word recognition and sentence processing in the working memory tasks are significant indicators of reading ability. Although the CET has recently incorporated fast reading tasks, the measurement approach, i.e., reading texts with multiple-choice items, is the same as the careful reading tasks. The present study has shown that word recognition processing and sentence processing efficiency could be related to reading rate. Although further factor analysis studies should be

conducted to examine the components of fast reading assessment, the DMDX-programmed word recognition processing and sentence processing efficiency tasks used in the present study would be strong candidates. Furthermore, the convenient and accurate DMDX computer program deserves serious consideration for the measurement of processing efficiency. This program yields information about participants' reaction time to each item, mean reaction time, standard deviation of reaction time, and correct response rate. It also provides the standard deviation of reaction time divided by the mean reaction time, which is termed as the coefficient of variation of reaction time by Segalowitz & Segalowitz (1993) and is used as the index of processing automaticity. The DMDX program also produces information about the test items, such as their mean reaction time and error rates, which would contribute to item screening and test development. In conclusion, word recognition and sentence processing tasks should be considered as candidates for reading rate measurement. The rich information generated by the DMDX program would make it a strong measurement tool for the measurement of processing efficiency.

Still another implication for reading assessment concerns the finding of the trade-off between sentence processing reaction time and recall. As shown in Figure 4.14, there exists a significant negative correlation between sentence processing reaction time and recall in the measurement of working memory capacity. This result is in accordance with the theoretical analysis of the competition for limited cognitive resources in comprehension (e.g., Baddeley & Hitch, 1974; Baddeley, 1986, 2007; Daneman &

Carpenter, 1980). There is an urgent need to help language teachers and students understand the nature of working memory, especially the trade-off between processing and memory, to motivate L2 learners to improve their processing efficiency. To echo Grabe's (2000) suggestion, it is advisable to design some working memory tasks in reading assessment, especially in tests utilized for the purposes of reading instruction and diagnosis.

In conclusion, this section has focused on three aspects of implications for L2 reading assessment. First, the findings of the present study have revealed the appropriateness of incorporating discourse knowledge measurement in reading assessment. Second, the results of this study indicate the possible contributions of word recognition and sentence processing tasks in the measurement of reading rate. Furthermore, the researcher determined that the DMDX program would serve as a powerful tool in the reading rate measurement. Finally, the findings of the current study have demonstrated the trade-off nature between processing and memorizing in working memory. Tasks could be designed to push L2 learners' working memory capacity, which could also be incorporated into instructional and diagnostic-oriented L2 reading assessments.

5.7 A DISCUSSION OF THE INSTRUMENTS FOR MEASURING WORD RECOGNITION AND WORKING MEMORY

In the present study, the instruments for measuring word recognition and working memory have been designed more as power tests than as speed tests. Power tests are

characterized by a smaller number of questions that are usually difficult to answer, while speed tests usually feature a larger number of easy questions, although a clear division between these two types of tests does not exist. Processing tasks are normally designed as a speed test. However, no research has focused on the exploration of a clear guideline for the cutoff criteria for correct response rates. In the present study, the word recognition measurement consists of 25 items, and the mean of correct response rates is 70.27% (see Table 4.1). The working memory measurement is composed of 28 items, and the mean of correct response rates is 78.61% (see Table 4.2). After correct response rates that are lower than 60% have been excluded, as shown in Table 4.12 the new means for word recognition and working memory are 74.95% and 79.99%, respectively. Whether the cutoff values of correct response rates are appropriate deserves further discussion.

Despite the consensus that processing tasks should measure response speed rather than response power, variance in response power, which is indicated by correct response rates, exists in almost every study. How to deal with the variance in response power remains an issue. Researchers have normally set a criterion based on an overall evaluation of participants' performances, the number of participants, and the number of items, which result in various cutoff values. For example, Van Gelderen et al. (2004) regarded correct response rates at 62.5% or lower on the word recognition and sentence verifications tests as missing data. In Conway et al.'s (2002) reading span test, 22 out of the 60 sentences for recall tasks were selected for comprehension measure. Participants who missed more than 10 of the 22 (45.5%) comprehension questions were removed

from the final analysis. In another working memory study, for which Conway was one of the researchers, Engle et al. (1999) set the cutoff value of the correct response rate of the reading span test at 85%.

In conclusion, the question of how to control the levels of item difficulty in the speed tests to measure processing efficiency warrants specific studies. Without clear guidelines, researchers have employed a variety of standards. The divergence of the cutoff values as to the correct response rates makes inferences drawn from research findings less comparable and less meaningful.

5.8 LIMITATIONS

Despite the strengths of the present study, three limitations have to be acknowledged. First, the reliabilities of the word recognition and working memory measurements are not high enough at .63 and .50, respectively. Although in theory the effect of low reliability could be corrected by the use of latent variables, word recognition and working memory are not treated as latent variables in this study. In future studies, a larger pilot study should be conducted to ensure higher qualities of items. Increasing the number of items would contribute to a more reliable measurement. The word recognition measurement consists of 25 items in the present study. The total test time would not be influenced much if 10 or 15 more items were added. For the working memory measurement, an increase in the item number is definitely necessary in future studies. The working memory measurement is composed of 28 sentences, which is divided into two parallel sets of two-, three-, four-, and five-sentence levels. Working memory studies

usually design three to five sets at each sentence group level. For example, Conway et al.'s (2002) reading span measurement had 60 sentences, which is in contrast to the 28 sentences in the present study.

Another limitation involves the scoring method of the working memory measurement. In the present study, an absolute span scoring method was used for the reading span test. As explained in Chapter three, the highest possible score was five, and lowest was zero. Without a large range, the scores are not sensitive enough to reflect the variance of recall among participants (Juffs, 2011; Oberauer & Suß, 2000). Conway et al. (2005) compared four different scoring methods — partial-credit unit scoring, all-or-nothing unit scoring, partial-credit load scoring, and all-or-nothing load scoring — and found that partial-credit scoring has an advantage over absolute span scoring (all-or-nothing scoring). The contrast between partial-credit and all-or-nothing scoring refers to whether points were awarded to the correctly recalled words, even though not all of the words in a set have been recalled. The contrast between unit and load scoring indicates whether the weight of items should be considered when scoring. Recalling a word from a five-sentence serial is supposed to be more difficult than recalling a word from a two-sentence serial. An example is if a participant has correctly recalled 3, 2, and 4 words (irrespective of order within a set) of the three sets of four-sentence serials. The scores yielded from the four methods are as follows:

Partial-credit unit scoring: $3/4 + 2/4 + 4/4 = 2.25$

All-or-nothing unit scoring: $0 + 0 + 1 = 1$

Partial-credit load scoring: 3 + 2 + 4 = 9

All-or-nothing load scoring: 0 + 0 + 4 = 4

When compared with other measures of working memory, Conway et al. found that partial-credit scores yielded higher internal consistency than all-or-nothing scores, and unit-weighted scoring had a slight advantage over load-weighted scoring. In conclusion, the measurement of working memory could have been more accurate if the partial-credit scoring method had been adopted.

Finally, the research design might have been more robust if certain information such as the topics of reading passages, the scores of fast reading items, and the reliability of the reading section has been accessible. Without information about the CET reading passages, participants' background knowledge could not be estimated. Consequently, the impact of background knowledge on reading ability was not examined. In the same vein, due of the lack of information about the scores for the fast reading tasks, an analysis of the relationship between the speed component of reading and the scores for fast reading was impossible. Owing to the lack of information about the reliability of the reading section that the participants attended, the researcher of this study had to utilize the information published in Yang & Weir (1998) to estimate the error variance of the scores on the CET reading section.

5.9 FUTURE RESEARCH

Some directions for future research have been implied in the previous sections in this chapter. This section aims to clarify the studies that the researcherplans to carry out

and make suggestions with respect to studies that scholars in related domains may conduct.

First, a study on the scoring method of the recall task in working memory should be conducted. Conway et al.'s (2005) conclusion that partial-credit scoring has an advantage over the absolute span was based on fictional data. Another study with real data would provide more evidence for the scoring guidelines of reading span tests.

Second, studies on the relationship between word recognition and reading speed should be designed. Although a strong relation between word recognition and reading ability has not been found in the present study, it is very likely that word recognition efficiency correlates strongly with reading speed (e.g., Haynes & Carr, 1990). Such a finding would greatly influence L2 reading instruction if fast reading tasks were increased on the reading section. Teachers would be more likely to realize the importance of processing efficiency and design instruction tasks, such as vocabulary games and extensive reading, to build up students' word recognition automaticity.

Third, scholars in listening, translation, and writing could conduct studies on the theory-based validity of the individual sections of the CET. Validation studies on the quality of the CET are currently insufficient in relation to its large population of stakeholders and impact on Chinese society.

Finally, teachers particularly could design studies on the concurrent validity of the CET, either on individual sections or the entire test. As revealed from the literature review in Chapter two, concurrent validity evidence for the CET is nil except for the sole

study by Yang & Weir (1998). Teachers may compare students' performance in class and on the school examinations with their performance on the CET. Teachers may also employ qualitative research methods to explore the differences between students' performances in class and on the CET if they identify some discrepancy. In conclusion, teachers and scholars outside of the CET committee could take a more active role in the CET validation studies.

5.10 CONCLUSION

With the aim to explore the construct validity of the CET reading section, this dissertation has embarked upon the modeling of reading ability. Six components have been chosen as observed variables of the latent variable of reading ability, namely, word recognition efficiency, working memory, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive reading skills. A pseudowords identification task programmed by DMDX, a revised version of Daneman & Carpenter's (1980) sentence reading span working memory test, Meara & Milton's (2002) Yes/No vocabulary tests, the test of syntactic knowledge used in Shiotsu & Weir's (2007) study, Abeywickrama's (2007) discourse knowledge test, and a revised version of Phakiti's (2008) strategy use questionnaire are utilized to measure the six observed variables. The results of confirmatory factor analysis show that the nine-observed-variable, one-factor confirmatory model was superior to the other three competing models. The nine indicators of reading ability are word recognition reaction time, word recognition correct response rates, working memory reaction time, working memory correct response rates,

recall, semantic knowledge, syntactic knowledge, discourse knowledge, and metacognitive strategy.

With the baseline confirmatory factor model of reading ability as well as participants' scores on the CET reading section, a structural model has been analyzed. The results indicated that the structural model showed good model fit indices. The path from reading ability to test performance on the CET reading section was .75, which implied that participants' test performance and scores on the CET reading section were strongly underlined by their actual reading ability. Moreover, 56.5% of the variance of test performance could be explained by reading ability, which is a large portion considering that the CET reading section accounted for only 35% of the entire CET. Therefore, the path value from reading ability to test performance revealed in the present study was strong enough to provide positive evidence for the construct validity of the CET reading section. The scores on the CET reading section were justifiable to a large degree for use in drawing inferences about participants' reading ability.

Appendix A — Participant recruitment flyer (Chinese version)

欢迎参加一项关于英语阅读的研究

不论你的英语阅读能力高低，只要你参加了2010年12月18日的全国英语四级考试，本人真诚地邀请你参加这项研究。它含有一个问卷调查和五个与英语阅读有关的分项测试，大约需要90分钟我们将收集你的CET阅读部分的成绩但不会单独报道。完成各部分的测试，你现场即可得到现金30元以补偿你所付出的时间。 你的信息将严格保密，测试结果不会影响你的任何考试成绩。

所需时间：	90分钟
酬劳：	30 元
参加者条件:	参加了2010年12月18日的全国英语四级考试 （请带准考证）
时间：	8:00 am - 6:00 pm （欢迎通过电话或邮件预约）
日期：	2011年 1月24日至2011 年2月20日
地点:	枫园外语学院三楼3036室
联系人：	陈静
联系电话：	18963995450; 13207110728; 13429899434;
邮箱地址：	reading.research@yahoo.com.cn

Appendix B — Participant recruitment flyer (English version)

Invitation to Participate in a Study on English Reading

You are invited to participate in a study on English reading if you attended the CET-4 on December 18th, 2010, regardless of your English proficiency. The participation consists of completing a questionnaire, three paper-pencil tests, and two computer-based tests. Your CET reading scores will be collected but will not be reported or studied individually. The total amount of time needed is about 90 minutes. You'll be compensated for your time with 30 RMB cash upon completion of all research activities. The result will not influence scores on any other tests and will be kept confidential.

Amount of time: 90 minutes

Compensation: 30 RMB

Requirement of participants: Attended the CET-4 on December 18th, 2010; please show your admission card

Time: 8:00 am – 8:00 pm （please make appointment via phone calls or email）

Date: January 24th - February 20th, 2011

Location: Room 3036, School of Foreign Languages and Literature

Coordinator: Jing Chen

Phones: 13207110728； 13429899434； 18963995450

E-mail: reading.research@yahoo.com.cn

Appendix C — Personal information sheet (Chinese version)

请回答下列八个关于你的背景和英语学习的问题。 你不愿意提供的信息或不想回答的问题可以空着。 谢谢你的合作!

1. 专业：_____

2. 四级准考证号码: _____

3. 性别： _____男 _____女

4. 年龄： _____岁

5. 你总共学了多少年英语？_____ 年。

6. 你从几岁开始学习英语（每星期至少一个小时）？ _____ 岁。

7. 在大学期间，除了上英语课而外，你每星期花几个小时学习和练习英语？_____ 小时。

8. 你认为英语对你将来的工作和学习很重要吗？

A. 很重要
B. 重要
C. 不清楚
D. 不重要

Appendix D — Personal information sheet (English version)

Please answer the following eight questions about your background and about your English learning experience. You may leave any question unanswered if you do not want to present the information. Thank you for your cooperation!

1. Major: _____

2. The CET admission card number: _____

3. Gender: _____Male _____Female

4. Age _____

5. How many **years** have you studied English (Including the years in college)?

_____ Years.

6. Age at which you first began to study English: _____

7. How many **hours per week** do you study or practice English **outside** of your English class? ____hrs/week

8. How important do you think English will be to your future job?

A. very important

B. important

C. I don't know

D. not important

Appendix E — Measurement of word recognition

This is the input script for the DMDX computer software to measure word recognition. It is written in Word and saved in rich text format (.rtf).

```
<n 30><cr><nfb><fd 200><t 3500><id "keyboard"><vm 1024,768,768,16,60>
! This is a word recognition task;
! Item number AB;
! A=answer (+=right one is the correct answer, -=left one is the correct answer);
! B=Trial number;
00<ln -3> “This is a word recognition test.”,
<ln -1> “You’ll see a pair of made-up words each time, e.g. kake dake.”,
  <ln 1> “Pronounce them out according to rules.”,
  <ln 3> “One of them sounds like a real word, e.g. kake sounds like cake.”,
<ln 5> “Press SPACEBAR to continue.”;
00<ln -1> “Press the left SHIFT key if the left one sounds like a real word.”,
<ln 1> “Press the right SHIFT key if the right one sounds like a real word.”,
<ln 3> “Press SPACEBAR to continue.”;
00 <ln -3> “Let’s practice three items first.”,
  <ln -1> “Respond as quickly and accurately as you can.”,
<ln 1> “Ready? Press SPACEBAR to begin.”;
-100 <% 30> “+ +”/*“kake   dake”/;
+200 <% 30> “+ +”/*“warld werld”/;
+300 <% 30> “+ +”/*“threa threi”/;
00 “End of practice. Press SPACEBAR to start the real task.”;
+1 <% 50> “+ +”/*“filst       ferst”/;
-2<% 50> “+ +”/*“bote        boaf”/;
+3 <% 50> “+ +”/*“broave   braive”/;
-4<% 50> “+ +”/*“lurn        lurm”/;
```

```
+5 <% 50> “+ +”/*“kleeze   pleeze”/;
-6<% 50> “+ +”/*“fite      fipe”/;
+7<% 50> “+ +”/*“neach     teech”/;
+8<% 50> “+ +”/*“threp     thrue”/;
-9<% 50> “+ +”/*“tirn      turt”/;
+10<% 50> “+ +”/*“glog     klok”/;
-11<% 50> “+ +”/*“katch    gatch”/;
-12<% 50> “+ +”/*“craul    crail”/;
+13<% 50> “+ +”/*“meave          leeve”/;
-14<% 50> “+ +”/*“teetch   neetch”/;
+15<% 50> “+ +”/*“phleer   phloar”/;
+16<% 50> “+ +”/*“hote     hoap”/;
-17<% 50> “+ +”/*“plaice   plice”/;
+18<% 50> “+ +”/*“rheatsh rheetch”/;
+19<% 50> “+ +”/*“tane     rane”/;
+20<% 50> “+ +”/*“plime    klime”/;
-21<% 50> “+ +”/*“wate     wame”/;
+22<% 50> “+ +”/*“symck   synck”/;
-23<% 50> “+ +”/*“naimb    baimb”/;
-24<% 50> “+ +”/*“knine    knime”/;
+25<% 30> “+ +”/*“phean   phaim”/;
00 “End of task. Thank you for participation.”;
```

Appendix F — Measurement of working memory

This is the input script for the DMDX computer software to measure working memory. It is written in Word and saved in rich text format (.rtf).

<n 28><cr><nfb><fd 200><t 30000><id "keyboard"><vm 1024,768,768,16,60>
! This is a working memory task;
! Item number AB;
! A=answer (+=the statement is correct, -=the statement is incorrect);
! B=Trial number;
00 <ln -1> “This is a working memory task.”,
<ln 1> “Read each sentence and memorize the attached red word.”,
<ln 3> “e.g., Two plus two makes ten. warm.”,
<ln 5> “Press SPACEBAR to continue.”;
00<ln -1> “Press YES if the statement is correct.”,
<ln 1> “Press NO if the statement is incorrect.”,
<ln 3> “Press SPACEBAR to continue.”;
00 <ln -1> “Respond as quickly and accurately as you can.”,
<ln 1> “Press SPACEBAR to receive a 2-sentence practice test.”;
-100 <% 60> “+ +”/*“Two plus two makes five. Saturday”/;
+200 <% 60> “+ +”/*“Winter is cold in Beijing. candy”/;
00 <ln -1> “Write down the red word in each sentence”,
<ln 1> “in the order of presentation on the answer sheet.”,
<ln 3> “Press SPACEBAR to continue.”;
00 “Ready? Press SPACEBAR to receive a 2-sentence real test.”;
+1 <% 60> “+ +”/*“The moon moves around the earth. water”/;
-2 <% 60> “+ +”/*“Wednesday is the last day of a week. fast”/;
00 <ln -1> “Write down the red word in each sentence”,
<ln 1> “in the order of presentation on the answer sheet.”,

<ln 3> “Press SPACEBAR to receive another 2-sentence level test.”;
-3 <% 60> “+ +”/*“Water is a kind of gas. man”/;
+4 <% 60> “+ +”/*“Five plus five makes ten. telephone”/;
00 <ln -1> “Write down the red word in each sentence”,
<ln 1> “in the order of presentation on the answer sheet.”,
<ln 3> “Press SPACEBAR to receive a 3-sentence level test.”;
-5 <% 60> “+ +”/*“Obama is the president of France. study”/;
+6 <% 60> “+ +”/*“A cat is smaller than an elephant. tree”/;
+7 <% 60> “+ +”/*“A grandfather is the mother of a person’s father. computer”/;
00 <ln -1> “Write down the red word in each sentence”,
<ln 1> “in the order of presentation on the answer sheet.”,
<ln 3> “Press SPACEBAR to receive another 3-sentence level test.”;
+8 <% 60> “+ +”/*“Bees produce honey. city”/;
-9 <% 60> “+ +”/*“Human beings can live without water. library”/;
-10<% 60> “+ +”/*“China is east of Japan. bird”/;
00 <ln -1> “Write down the red word in each sentence”,
<ln 1> “in the order of presentation on the answer sheet.”,
<ln 3> “Press SPACEBAR to receive a 4-sentence level test.”;
+11 <% 60> “+ +”/*“Most people don’t work on Sundays. university”/;
-12 <% 60> “+ +”/*“A basket ball is in the shape of square. cry”/;
-13<% 60> “+ +”/*“A table usually has seven legs. face”/;
+14<% 60> “+ +”/*“A year is made of 12 months. card”/;
00 <ln -1> “Write down the red word in each sentence”,
<ln 1> “in the order of presentation on the answer sheet.”,
<ln 3> “Press SPACEBAR to receive another 4-sentence level test.”;
+15<% 60> “+ +”/*“Two times two makes four. morning”/;
-16<% 60> “+ +”/*“A car runs faster than a train. river”/;
+17<% 60> “+ +”/*“Red and yellow makes orange. work”/;

-18<% 60> “+ +”/*“Spring is the hottest season. bottle”/;

00 <ln -1> “Write down the red word in each sentence”,

<ln 1> “in the order of presentation on the answer sheet.”,

<ln 3> “Press SPACEBAR to receive a 5-sentence level test.”;

-19<% 60> “+ +”/*“A professor’s major work is cooking. brother”/;

+20<% 60> “+ +”/*“A football is bigger than a baseball. cat”/;

+21<% 60> “+ +”/*“All rivers flow to the ocean. grass”/;

-22<% 60> “+ +”/*“China has a history of only 500 years. school”/;

-23<% 60> “+ +”/*“Ten minus four makes five. holiday”/;

00 <ln -1> “Write down the red word in each sentence”,

<ln 1> “in the order of presentation on the answer sheet.”,

<ln 3> “Press SPACEBAR to receive another 5-sentence level test.”;

+24<% 60> “+ +”/*“Plants need sunlight. dance”/;

-25<% 60> “+ +”/*“Human beings have smaller brains than animals. apple”/;

-26<% 60> “+ +”/*“Yellow River is in India. bicycle”/;

+27<% 60> “+ +”/*“Paper is made from trees. fight”/;

+28<% 60> “+ +”/*“December is the last month of a year. great”/;

00 “End of the task. Thank you!”;

Appendix G — Measurement of semantic knowledge

Instructions*: There are four sets in this test, with 60 words in each set. Some of the words are not real English words. Put a check mark (√) in the square if you know the meaning of the word.* ***Leave the square blank*** *if you do not know the meaning of the word or if it is not a real English word.*

Set one

1 ☐ adair	2 ☐ gumm	3 ☐ cliff	4 ☐ stream
5 ☐ system	6 ☐ position	7 ☐ law	8 ☐ whaley
9 ☐ contrivial	10 ☐ pocock	11 ☐ amuse	12 ☐ museum
13 ☐ turn over	14 ☐ prefer	15 ☐ method	16 ☐ generous
17 ☐ hoult	18 ☐ organize	19 ☐ normal	20 ☐ everywhere
21 ☐ knowledge	22 ☐ relation	23 ☐ whitrow	24 ☐ director
25 ☐ criminal	26 ☐ snell	27 ☐ check in	28 ☐ useful
29 ☐ enter	30 ☐ berrow	31 ☐ though	32 ☐ sale
33 ☐ cage	34 ☐ limidate	35 ☐ handkerchief	36 ☐ pernicate
37 ☐ sight	38 ☐ humberoid	39 ☐ pring	40 ☐ fountain
41 ☐ eldred	42 ☐ reward	43 ☐ eluctant	44 ☐ guess
45 ☐ persuade	46 ☐ hubbard	47 ☐ stace	48 ☐ aim
49 ☐ detailoring	50 ☐ stimulcrate	51 ☐ aunt	52 ☐ bend
53 ☐ deny	54 ☐ bastionate	55 ☐ shot	56 ☐ maker
57 ☐ rabbit	58 ☐ steady	59 ☐ weekly	60 ☐ inform

Set two

1 ☐ sandy	2 ☐suddery	3 ☐military	4 ☐ interval
5 ☐ overcoat	6 ☐overcome	7 ☐ get out of	8 ☐ structure
9 ☐ typist	10☐ break off	11☐ heap	12 ☐ majority
13 ☐ remedy	14 ☐ cure	15☐ acklon	16 ☐ jarvis
17 ☐ plus	18 ☐ duffin	19☐ accuse	20 ☐ impress
21 ☐ twose	22 ☐ oestrogeny	23☐ provision	24 ☐ recenticle
25 ☐ fluctual	26 ☐ feel up to	27☐ wipe out	28 ☐ staircase
29 ☐ cambule	30☐ ridout	31☐ kind-hearted	32 ☐ border
33 ☐ dozen	34 ☐ mystery	35 ☐ apartment	36 ☐ wilding
37 ☐ condimented	38 ☐ theory	39 ☐ leave out	40 ☐ puzzle
41 ☐ charactal	42 ☐ emphasise	43 ☐ send in	44 ☐ check over
45 ☐ wray	46 ☐ hapgood	47 ☐ tend	48 ☐ escrotal
49 ☐ grip	50 ☐ catch up with	51 ☐ cut off	52 ☐ urge
53 ☐ menstruable	54 ☐ batcock	55 ☐ vital	56 ☐ moffat
57 ☐ complicate	58 ☐ smack	59 ☐ exist	60 ☐ semaphrodite

Set three

1 ☐ lessen	2 ☐ oak	3 ☐ mosquito	4 ☐ litholect
5 ☐ quorant	6 ☐ proceed	7 ☐ interfere	8 ☐ put up with
9 ☐ algebra	10☐ scurrilize	11☐ cottonwool	12☐ lobby
13☐ give away	14☐ trudgeon	15☐ bodelate	16☐ tighten
17☐ shady	18☐ bance	19☐ awkward	20☐ wartime
21☐ draconite	22☐ folksong	23☐ outskirts	24☐ technology
25☐ stand in for	26☐ victory	27☐ antique	28☐ chart

29☐ rot	30☐ manly	31☐ compose	32☐ risk
33☐ pea	34☐ tunnel	35☐ justal	36☐ call up
37☐ combustulate	38☐ democracy	39☐ opie	40☐ scudamore
41☐ homoglyph	42☐ abrogative	43☐ react	44 ☐ haque
45☐ nickling	46☐ bench	47☐ snack-bar	48☐ charlett
49☐ harden	50☐ scorn	51☐ equality	52☐ jewel
53☐ pass away	54☐ webbert	55☐ kiley	56☐ woolnough
57☐ hijack	58☐ baldock	59☐ farther	60☐ dose

Set four

1 ☐ ambiguous	2 ☐ prosecute	3 ☐ harness	4 ☐ allimer
5 ☐ constrain	6 ☐ blagegerage	7 ☐ referral	8 ☐ gospel
9 ☐ cloakery	10 ☐ accessory	11 ☐ illuminate	12 ☐ bait
13 ☐ pedestrian	14 ☐ cupoid	15 ☐ ion	16 ☐ conversely
17 ☐ fallity	18 ☐ articulate	19 ☐ disguise	20 ☐ verge
21 ☐ floratious	22 ☐ hite	23 ☐ deploy	24 ☐ counselor
25 ☐ eternal	26 ☐ tread	27 ☐ interfate	28 ☐ inference
29 ☐ mensible	30 ☐ denote	31 ☐ sway	32 ☐ ample
33 ☐ binary	34 ☐ symmetry	35 ☐ murray	36 ☐ inhibition
37 ☐ orrade	38 ☐ mandatory	39 ☐ retrieval	40 ☐ psychic
41 ☐ wallage	42 ☐ entrant	43 ☐ equitable	44 ☐ prophet
45 ☐ pharicise	46 ☐ hurdle	47 ☐ multitude	48 ☐ convolition
49 ☐ quirty	50 ☐ tactical	51 ☐ treggle	52 ☐ vessy
53 ☐ higher order	54 ☐ watler	55 ☐ emit	56 ☐ surman
57 ☐ sprinkle	58 ☐ trimble	59 ☐ contamination	60 ☐ endorsement

Appendix H — Measurement of syntactic knowledge

Instructions: *There are 32 incomplete sentences in this part. For each sentence there are four choices marked A), B), C) and D). Choose the ONE answer that best completes the sentence.*

1. ____ of the students has started the course.

A. Several B. Both C. Neither D. Most

2. The metal was ____ hot that he couldn't touch it.

A. very B. too C. so D. extremely

3. By the time this course finishes ____ a lot about engineering.

A. I will learn B. I learn

C. I will have learnt D. I have learnt

4. _____ many years he studied hard for his doctorate.

A. During B. For C. Since D. From

5. We found ____ to understand his lecture.

A. difficulty B. difficult C. so difficult D. it difficult

6. My research findings were not ____ to be published.

A. interesting so B. interesting enough

C. enough interesting D. so interesting

7. As a result of his lectures she ____ by this new approach to teaching.

A. was influenced B. has influenced

C. influenced D. had influenced

8. If he had known the problem, he ____ the task.

A. will not have undertaken B. had not undertaken

C. should not undertake D. would not have undertaken

9. ____ a pity you did not check the figures with your partner.

A. What's B. That's C. There's D. It’s

10. The penguin is a bird adapted to life ____ on land and in water.

A. both B. not only C. and D. either

11. My results are the same ____ yours.

A. that B. as C. than D. like

12. Caramel is a brown substance ____ by the action of heat on sugar.

A. form B. forming C. formed D. forms

13. I ____ to finish my thesis next year.

A. intend B. think C. decide D. will

14. You'd better ____ to the doctor next time you feel ill.

A. to go B. going C. go D. gone

15. ____ I need is a long holiday.

A. What B. That C. Which D. The which

16. He is ____ proud man that he would rather fail than ask for help.

A. so a B. such C. a so D. such a

17. Your English is very good. "

"It should be. I ____ it ever since I started school. "

A. have been learning B. was learning

C. had learned D. had been learning

18. If only he ___ down the results when he did the experiments!

A. writes B. had written C. has written D. was writing

19. Vitamin C, discovered in 1932, ___ first vitamin for which the molecular structure was established.

A. the B. was the C. as the D. being the

20. The behavior of gases is explained by ____ the kinetic theory.

A. what scientists call B. what do scientists call

C. scientists they call D. scientists call it

21. Ironically, sails were the salvation of many steamships ___ mechanical failures.

A. they suffered B. suffer

C. were suffered D. that had suffered

22. ____ some mammals came to live in the sea is not known.

A. Which B. Since C. Although D. How

23. ____ their nests well, but also build them well.

A. Not only brown thrashers protect

B. Protect not only brown thrashers

C. Brown thrashers not only protect

D. Not only protect brown thrashers

24. The name Nebraska comes from the Oto Indian word 'nebrathka', ____ flat water.

A. to mean B. meaning C. it means D. by meaning

25. Biochemists use fireflies to study bioluminescence, ____.

A. the heatless light given off by certain plants and animals

B. certain plants and animals give off the heatless light

C. which certain plants and animals give off the heatless light

D. is the heatless light given off by certain plants and animals

26. Rich tobacco and champion race horses have ____ of Kentucky.

A. long been symbols B. been long symbols

C. symbols been long D. long symbols been

27. Today's libraries differ greatly from ____ .

A. the past B. those of the past C. that are past D. those past

28. Conifers first appeared on the Earth ____ the early Permian period, some 270 million years ago.

A. when B. or C. and D. during

29. There are very few areas in the world ____ be grown successfully.

A. where apricots can B. apricots can

C. apricots that can D. where can apricots

30. ____ 'a baby turtle is hatched, it must be able to fend for itself.

A. Not sooner than B. No sooner

C. So soon that D. As soon as

31. Tungsten, a gray metal with the ____ is used to form the wires in electric light bulbs.

A. point at which it melts is the highest of any metal

B. melting point is the highest of any metal

C. highest melting point of any metal

D. metal's highest melting point of any

32. Rattan comes from ____ of different kinds of palms.

A. its reedy stems B. the reedy stems

C. the stems are reedy D. stems that are reedy

Appendix I — Measurement of discourse knowledge

Instructions: *There are two tasks in this part. Task A requires you to read the passage titled "The Development of Learning Theory" at your normal speed and then fill in each blank with one word to make the passage coherent in meaning. Task B consists of six mini passages. At the end of each passage, there is a question about the overall structure of each passage.*

Task A Cloze

The Development of Learning Theory

Speculation about the mental aspects of human beings goes back to the Greek philosophers. A more specific interest in learning was well established by the 1600s, when it found expression in the writings of John Locke, David Hume, and other British philosophers. Actual experiments on human memory, 1)_____ did not start until the 1880s. The German psychologist, Hermann Ebbinghaus, conducted elaborate 2)_____ on himself as a subject 3)_____ learned and tried to recall "nonsense syllables". He showed the effects on learning of 4) _____ independent variables as length of the materials 5) _____ the number of repetitions in presenting it to a subject. In studying how memory lapsed, 6) _____ formulated his famous "curve of forgetting", showing the 7)_____ between time and accuracy of 8)_____. Such early studies were innovative and thorough, 9)_____ the use of human beings in experiments had many 10)_____. Willing subjects were hard to obtain, control 11)_____ extraneous conditions was difficult, 12)_____human beings could be subjected to only a 13)_____range of experimental conditions.

Focusing on animals offered a way to study learning that 14)_____ these dangers. We noted 15)_____ a conviction had increased in comparative psychology that processes characterizing animals would 16)_____ be found in human beings and vice versa. During the first decade of the 1900s this phylogenetic

perspective focused the 17)_____ of psychologists on the study of 18)_____ among animals as a way to uncover basic and universal principles of the process. Just 19)_____ animal subjects could be used for medical 20)_____ that would yield conclusions 21)_____ to human beings, it seemed to the comparative psychologists 22)_____ studies of 23)_____ learning could provide the key to how human beings 24)_____ new forms of behavior. Animals were readily 25) _____, they introduced far 26)_____ extraneous conditions (such as language) that 27)_____ confound research, and 28)_____ could be used in experiments under 29)_____ that would be 30)_____ for human beings.

Task B Recognition of top-level organization

Instructions: *Read the following six passages at your normal speed. At the end of each passage, there is a question about the overall structure of each passage. Choose the one answer that best describes the overall organization of the passages.*

For example, you read

While mental health experts maintain that it's important to make friends in your new environment and be involved in the college community, it's equally crucial not to let bonds dissolve with the people you knew before college. They, after all, know you better than people you first met two months ago.

Then you will be asked to answer the following question about the overall structure of the passage.

The purpose of this passage is to _____.

A. Compare two kinds of opinions
B. Provide solutions to a problem
C. Explain the causation of a phenomenon
D. Present a claim

The correct answer is D. In the first sentence, the passage presents a claim, contacting old friends is equally important as making new friends. The second sentence reinforces the importance the old friends.

Passage one

Despite the argument that smoking is harmful, not everyone agrees. Certainly, smoking has been related to lung cancer, high blood pressure, and loss of appetite. But, for some people smoking relieves tension.

The purpose of this passage is to _____.

A. Compare two kinds of opinions
B. Provide solutions to a problem
C. Explain the causation of a phenomenon
D. Present evidence to a claim

Passage two

Did you know that people who get enough sleep (about 7-9 hours a night) are more likely to have higher productivity, feel more energetic throughout the day, and experience less stress? Sleep is crucial for concentration, memory formation, and repairing and rejuvenating the cells of the body. Both mentally and physically, a good night's sleep is essential for your health and your energy.

The purpose of this passage is to _____.

A. Compare two kinds of opinions
B. Provide solutions to a problem
C. Explain the causation of a problem
D. Present a claim

Passage three

Cirrus clouds are thin and delicate, whereas cumulus clouds look like cotton balls. Nimbus clouds are dark and ragged, and stratus clouds appear dull in color and cover the entire sky.

Overall, this passage is ____.

A. A collection of descriptions

B. A presentation of a method
C. A description of the sequence of actions
D. A presentation of a claim

Passage four

Traditionally, America's fast-food companies have hired teenagers. While teenagers provide cheap labor, they are sometimes unreliable. Consequently, fast-food companies are looking into another source of cheap labor — the elderly. Older people are less likely to skip a day of work or quit without giving notice.

The passage _____.

A. Compares two arguments
B. Presents an argument
C. Presents a solution to a problem
D. Presents evidence to a claim

Passage five

Many of the incoming members of Congress campaigned for reining in federal spending and getting the budget balanced. Critics argue that cutting expenditures now could threaten the economic recovery, while advocates say that excessive deficits stall job growth.

The purpose of this passage is to _____.

A. Present two arguments
B. Present evidence to a claim
C. Provide solution to a problem
D. Describe a phenomenon

Passage six

Since 1782 the bald eagle has been the national emblem of the United States. At that time bald eagles nested throughout most of North America. In the late 1960s bald eagles had almost disappeared from the eastern United States. Now they can be found again, especially in the Great Lakes Region, around Chesapeake Bay, and in Maine and Florida**.**

This passage is arranged according to

A. The sequence of what has happened
B. The degree of importance
C. Difference aspects of an animal's characteristics
D. Degree of the author's emotion

Appendix J — Measurement of metacognitive strategy in reading

Instructions: *Read the following 22 statements and circle the letter that best describes what you normally do when you read English materials.*

1. I ask myself what my reading purposes are (*e.g.* to finish an assignment, to know about international news, to learn new words) before I begin to read.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
2. I scan through the reading materials before I read them carefully.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
3. I plan how many pages or how many minutes to read before I begin reading.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
4. I try to understand the main content without looking up every word.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
5. I try to understand the relationships between ideas in the text.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
6. I try to interpret the author's implied meaning.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
7. I summarize the main information in the text.
   A. Never B. Rarely C. Sometimes
   D. Often E. Usually F. Always
8. I anticipate what the author will write next while reading.
   A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

9. I relate the information from the text to my prior experience or to what I have learned before.

   A. Never B. Rarely C. Sometimes

   D. Often E. Usually F. Always

10. I apply my grammar knowledge when I do not understand difficult sentences.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

11. I guess meanings of unknown words using root words.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

12. I guess meanings of unknown words using context clues.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

13. I can differentiate which information is more or less important.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

14. I am aware of time limitations.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

15. I ask myself if I understand the text.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

16. I know when I lose concentration, or when I feel worried, tense, or unmotivated while reading.

    A. Never B. Rarely C. Sometimes

    D. Often E. Usually F. Always

17. I notice where I am confused in the text.

A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

18. I adjust reading rates according to different reading purposes or different materials.

A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

19. I reread the text several times when I feel I do not understand it.

A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

20. I try to correct my misunderstanding in reading tasks when found.

A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

21. I evaluate whether my reading goals have achieved after I finish reading.

A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

22. I evaluate the quality of the reading material (whether it is well written, whether it is helpful to my English learning, whether it is instructive) after I finish reading.

A. Never B. Rarely C. Sometimes

D. Often E. Usually F. Always

Appendix K — Consent form

Title: Exploring evidence for the construct validity of the reading comprehension section of the College English Test: A component skills approach

IRB protocol # 2010 - 11 - 0128

Conducted By: Min Gui

Of The University of Texas at Austin: *Department of Curriculum & Instruction, Foreign Language Education program / SZB528*; Telephone: 001-512-2324080; 001-512-4964553

You are being asked to participate in a research study. This form provides you with information about the study. The person in charge of this research will also describe this study to you and answer all of your questions. Please read the information below and ask any questions you might have before deciding whether or not to take part. Your participation is entirely voluntary. You can refuse to participate or stop participating at any time without penalty or loss of benefits to which you are otherwise entitled. You can stop your participation at any time and your refusal will not impact current or future relationships with UT Austin or Wuhan University. To do so simply tell the researcher you wish to stop participation. The researcher will provide you with a copy of this consent for your records.

The purpose of this study is to examine the skills that influence English reading ability and the relationship between reading ability and the scores on the CET reading section.

If you agree to be in this study, we will ask you to complete following tasks:

- One questionnaire about strategy use in English reading and background, about 5 minutes;
- One vocabulary test, about 15 minutes;
- One syntactic knowledge test, about 25 minutes;
- One discourse knowledge test, about 30 minutes;

- One word recognition test on the computer, about 5 minutes; and
- One working memory test on computer with a sheet to write down words attached to sentences, about 10 minutes.

You will also be asked to allow the researcher to collecting your CET reading scores. The scores will not be reported individually and will not be connected with your name.
Total estimated time to participate in study is about 90 minutes.

Risks of being in the study

- The risk associated with this study is no greater than everyday life.
- There might be a possible risk for loss of confidentiality.
- The decision to participate or not will not affect your scores of any others tests.
- This study may involve risks that are currently unforeseeable. If you wish to discuss the information above or any other risks you may experience, you may ask questions now or call the Principal Investigator listed on the front page of this form.

Benefits of being in the study: There will be no direct benefit toward participants in the study. However, foreign language instruction and assessment might benefit from the research findings of this study.

Compensation: 30 *Yuan* if you complete all research activities.

Confidentiality and Privacy Protections:

- You will be required to provide your CET admission card. The card number will be coded after merging the test scores with the remaining data.
- The data resulting from your participation may be made available to other researchers in the future for research purposes not detailed within this consent form. In these cases, the data will contain no identifying information that could associate you with it, or with your participation in any study.
- The paper materials will be stored in a locked cabinet at the researcher's home.

The electronic data will be kept in the researcher's private and password protected computer. The original CET admission card numbers will be deleted after they are coded. All data will be labeled with a code and your name will not be used to identify the responses to the tests, questionnaires or CET reading scores.

- Authorized persons from The University of Texas at Austin, members of the Institutional Review Board have the legal right to review your research records and will protect the confidentiality of those records to the extent permitted by law. All publications will exclude any information that will make it possible to identify you as a subject. Throughout the study, the researchers will notify you of new information that may become available and that might affect your decision to remain in the study.

Contacts and Questions:

If you have any questions about the study please ask now. If you have questions later, want additional information, or wish to withdraw your participation call the researchers conducting the study. Their names, phone numbers, and e-mail addresses are at the top of this page.

If you would like to obtain information about the research study, have questions, concerns, complaints or wish to discuss problems about a research study with someone unaffiliated with the study, please contact the IRB Office at (512) 471-8871 or Jody Jensen, Ph.D., Chair, The University of Texas at Austin Institutional Review Board for the Protection of Human Subjects at (512) 232-2685. Anonymity, if desired, will be protected to the extent possible. As an alternative method of contact, an email may be sent to orsc@uts.cc.utexas.edu or a letter sent to IRB Administrator, P.O. Box 7426, Mail Code A 3200, Austin, TX 78713.

You will be given a copy of this information to keep for your records.

Statement of Consent:

I have read the above information and have sufficient information to make a decision about participating in this study. I consent to participate in the study.

Signature:_____ Date: _____

_____ Date: _____

Signature of Person Obtaining Consent

Signature of Investigator:_____ Date: _____

References

Abbott, M. L. (2006). ESL reading strategies: Differences in Arabic and Mandarin speaker test performance. *Language Learning, 56*(4), 633–670.

Abeywickrama, P.S. (2007). *Measuring the knowledge of textual cohesion and coherence in learners of English as a second language (ESL)*. Unpublished PhD Dissertation. University of California, Los Angeles.

Adams, M. J. (1994). Modeling the connections between word recognition and reading. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading,* 4th ed. (pp.830–863). Newark, DE: International Reading Association.

Akamatus, N. (2003). The effects of first language orthographic features on second language reading in text. *Language Learning, 53,* 207–231.

Alderson, J. C. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium* (pp. 203–219). Alexandria, VA: TESOL.

Alderson, J. C. (2000). *Assessing reading*. New York: Cambridge University Press.

Alexander, P. A., Graham, S., & Harris, K. R. (1998). A perspective on strategy research: Progress and prospects. *Educational Psychology Review, 10*(2), 129–154.

American Educational Research Association, American Psychological Association, & national Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association .

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal*, *75*, 460–472.

Anderson, J. R. (1995). Cognitive psychology and its implications (4th ed.). New York: W. H. Freeman.

Anderson, N. J, Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8(1),* 41–66.

Anderson, R. C., & Freebody, P. (1983). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research review* (pp. 77–117). Newark, DE: International Reading association.

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading/language research: A research annual* (pp. 231–256). Greenwich, CT: JAI Press.

Anderson, R. C., & Pearson, P.D. (1988). A schema-theoretic view of basic processes in reading comprehension. In P.L. Carrell, J. Devine & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 37−55). Cambridge: Cambridge University Press.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535–556.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17* (1), 1–42.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon.

Baddeley, A. D. (2007). *Working memory, thought, and action*. New York: Oxford University Press.

Baddeley, A. D. & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 47−87). New York: Academic Press.

Baker, L. & Brown A. L. (1984). Metacognitive skills and reading. In Pearson, P. D., editor, *Handbook of reading research* (pp. 353–394). New York: Longman.

Barnett, M. (1986). Syntactic and lexical/semantic skill in foreign language reading: Importance and interaction. *Modern Language Journal, 70,* 343–349.

Bartlett, B. (1978). Top-level structure as an organizational strategy for recall of classroom text. Ph.D. dissertation, Arizona State University.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M. & van de Velde, H. (2001). Examining the yes-no vocabulary: Some methodological issues in theory and practice. *Language Testing, 18*, 235–274.

Berman, R. A. (1984). Syntactic components of the foreign language reading process. In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 139–156). Harlow, UK: Longman.

Bell, L. C., & Perfetti, C. A. (1994). Reading skills: Some adult comparisons. *Journal of Educational Psychology, 86,* 244–255.

Bentler, P.M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Berman, (1984). Syntactic components of the foreign language reading process. In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 139–156). Harlow, UK: Longman.

Bernhardt, E. B. (1991). *Reading development in a second language*. Norwood, NJ: Ablex.

Bernhardt, E. B. (2000). Second-language reading as a case study of reading scholarship in the 20th century. In P.B. Mosenthal, M. L. Kamil, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 791–811). Mahwah, NJ: Erlbaum.

Block, E. (1986). The comprehension strategies of second language readers. *TESOL Quarterly, 20,* 463–496.

Brantmeier, C. (2002). Second language reading strategy research at the secondary and university levels: Variations, disparities, and generalizability, *The Reading Matrix, 2*(3). 1–14**.**

Brisbois, J. E. (1995). Connections between first- and second-language reading. *Journal of reading behavior, 27*(4), 565–584.

Brown, A. L. (1980). Metacognition development and reading. In Spiro, R. J. Bruce, B. B. & Brewer, W. F., editors, Theoretical issues in reading comprehension. Hillsdale, NJ: Erhbaum, 453–481.

Brown, T., & Haynes, M. (1985). Literacy background and reading development in a second language. In T. H. Carr (Ed.), *The development of reading skills* (pp.19–34). San Francisco, CA: Jossey-Bass.

Bruner, J. (1981). The pragmatics of acquisition. In W. Deutsch (Ed.), *The child's construction of language* (pp. 39−55). New York: Academic Press.

Carr, T. H., Brown, T. L., Vavrus, L. G., & Evans, M.A. (1990). Cognitive skills maps and cognitive skill profiles: Componential analysis of individual differences in children's reading efficiency. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 1–55). San Diego: Academic Press.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19,* 727–752.

Carrell, P. L. (1988). Some causes of text-boundedness and schema interference in ESL reading. In P.L. Carrell, J. Devine & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 101−113). Cambridge: Cambridge University Press.

Carrell, P. L. (1989). Metacognitive awareness and second language reading. *Modern Language Journal, 73*, 121–133.

Carroll, J. B. (1971). Development of native language skills beyond the early years. In C. Reed (Ed.), *The learning of language* (pp. 97–156.). New York: Appleton-Century-Crofts.

Carver, R. P. (1990). *Reading rate: A review of research and theory*. New York: Academic Press.

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies of Reading, 1* (1), 3–43.

Carver, R. P. (2000). *The cause of high and low achievement*. Mahwah, NJ: Erlbaum.

Chamot, A. U., O'Malley, J. M. (1994). Instructional approaches and teaching procedures. In K. Sangenberg-Urbschat & R. Pritchard (Eds.), *Kids come in all languages: Reading instruction for ESL students* (pp.82–107). Newark, DE: International Reading Association.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp.32–70). Cambridge: Cambridge University Press.

Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language.* New York: Routledge.

Clarke, M. (1979). Reading in Spanish and English: Evidence from adult ESL learners. *Language Learning, 29,* 121−150.

Cizek, G. J. (2008). Sources of validity evidence for educational and psychological tests. *Educational and psychological Measurement, 68* (3), 397–412.

Conway, A. R. A., Cowan, N., Bunting, M. Therriault, D., & Minkoff, S. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30* (2), 163 ‑ 183.

Conway, R. A., Kane, M. J., Bunting, M., Hambrick, D. Wilhelm, O., & R. Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review 12* (5), 769–786.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52* (4), 281–302.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Memory and Verbal Behavior, 19,* 450–456.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3,* 499–545.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & and Liskin-Gasparro, J. E. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper.* Princeton, NJ.: Educational Testing Service.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128* (3), 309–331.

Enright, M. K., Grabe, W., Koda, K., Mesenthal, P., Mulcahy-Ernt, P., & Schedl, M. (1997). *TOEFL® 2000 reading framework: A working paper*. Princeton, NJ: Educational Testing Services.

Favreau, M., Komoda, M. K., & Segalowitz, N. (1980). Second language reading: Implication of the word-superiority effect in skilled bilinguals. *Canadian Journal of Psychology, 34,* 377–391.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911.

Gee, J. P. (2001). Reading as situated language: A sociocognitive perspective. Journal of Adolescent & Adult Literacy, 44, 714−725.

Geva, E. (1983). Facilitating reading through flowcharting. *Reading Research Quarterly, 18,* 384–405.

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language Learning, 43,* 5–43.

Goodman, K. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist, 6,* 126−135.

Goodman, K. (1981). Letter to the editor. *Reading Research Quarterly, 16* (3), 477−478.

Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The relationships between phonological sensitivity, syntactic processing, and verbal working memory in the

reading performance of third-grade children. *Journal of Experimental Child Psychology, 63,* 563–582.

Gough, P. B. (1972). One second of reading. In F. J. Kavanagh, & I. G. Mattingly (Eds.), *Language by eye and by ear* (pp.331−358). Cambridge: MIT Press.

Gough, P. B., & Tunmer, W. (1986). Decoding, reading, and reading disability. RASE: Remedial and Speical Education, 7, 6–10.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25, 375–406.

Grabe, W. (2000). Reading research and its implications for reading assessment. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium.* University of Cambridge Local examinations syndicate: Cambridge University Press.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Grabe, W. & Stoller, F. (2002). *Teaching and researching reading*. New York: Longman.

Graesser, A., Gernsbacher, M. A., & Goldman, S. R. (2003). *Handbook of discourse processes.* Mahwah, N.J.: Lawrence Erlbaum.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hamada, M., & Koda, K. (2010). Phonological decoding in second language word-meaning inference. *Applied Linguistics, 31*, 513–531.

Haynes, M. & Carr. T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 375–421). San Diego: Academic Press.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23* (3), 370–401.

Heath, S. B. (1981). The functions and uses of literacy. *Journal of Communication, 30* (1), 123−133.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* New York: Cambridge University Press.

Henriksen, B. (1999). Three dimensions of vocabulary development. In M. Wesche & T. S. Paribakht (Eds.), *Incidental L2 vocabulary acquisition: Theory, current research, and instructional implications* [special issue]. *Studies in Second Language Acquisition, 21,* 303–317.

Hosenfeld, C. (1977). A preliminary investigation of the reading strategies of successful and nonsuccessful second language learners. *System, 5,* 110–123.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An interdisciplinary Journal, 2,* 127 – 160.

Hoover, W. A., & Tunmer, W. E. (1993). The components of reading. In G. B. Thompson, W. E. Tunmer & T. Nicholson (Eds.), *Reading acquisition processes.* Clevedon, PA:Multilingual Matters.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6* (1), 1–55.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. New York: Macmillan. (Republished: Cambridge, MA: MIT Press, 1968).

Huibregtse, I., Admiraal, W. and Meara, P. (2002). Scores on a yes–no vocabulary test: correction for guessing and response style. *Language Testing, 19,* 227–245.

Jin, Y., & Yang, H. (2006). The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum, 19* (1), 21–36.

Juffs, A. (2001). Psycholinguistically oriented second language research. *Annual Review of Applied Linguistics, 21,* 207–220.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44* (2), 137–166.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122–149.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112* (3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and practice, 21* (1), 31–35.

Kintch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Kintsch, W. (1988). *Comprehension: A framework for cognition.* New York: Cambridge University Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.)New York: The Guilford Press.

Kleiman, G. M. (1975). Speech recording in reading. *Journal of Verbal Learning and Verbal Behavior, 14*, 323–339.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge: Cambridge University Press.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal, 80* (4), 450–460.

Koda, K. (1988). Cognitive process in second language reading: Transfer of L1 reading skills and strategies. *Second Language Research, 4,* 133–156.

Kunnan, A. J. (Ed). (2000). Fairness and validation in language assessment: selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida. Cambridge, UK; New York, NY, USA: Cambridge University Press.

Kunnan, A. J. (Ed). (1998). Validation in language assessment: selected papers from the 17th Language Testing Research Colloquium, Long Beach, California. New Jersey: Mahwah.

Kwon, H. J. (2010). The Nature of metacognitive knowledge for reading comprehension strategy and language use by highly proficient learners of English. Unpublished doctoral dissertation, The University of Texas at Austin.

LaBerge, D., & Samuel, D. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323.

Laufer, B. & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*(1), 33–51.

Lee, J., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly, 31*, 713–739.

Levy, B. A. (1975). Vocalization and suppression effects in sentence memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 304–316.

Levy, B. A., Hinchley, J. (1990). Individual and developmental differences in the acquisition of reading skills. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 81–128). San Diego: Academic Press.

McNamara, T. F. (2007). Language assessment in foreign language education: The struggle over constructs. *Modern Language Journal, 91* (2), 280–282.

Meara, P. (1992). *EFL vocabulary tests.* Swansea: Centre for Applied Language Studies, University College Swansea.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge University Press.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4,* 142–45.

Meara, P. & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society*. CILT.

Meara, P., & Milton, J. L. (2002). *The Swansea vocabulary levels test.* Newbury: Express.

Messick, S. A. (1989). Validity. In Linn, R. L. (Ed) *Educational Measurement*, (3rd edition). New York: American Council on Education/ Macmillan Publishing Company.

Messick, S. (1992). Validity of test interpretation and use. In M.C. Alkin (ed.), *Encyclopedia of Educational Research* (6th edition). New York: Macmillan.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North-Holland.

Meyer, B. J. F. (1985). Prose analysis: Purposes, procedures and problems. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text* (pp. 269–304). Hillsdale, NJ: Lawrence Erlbaum.

Meyer, B. J. F., & Poon, L. W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology, 93*, 141–159.

Meyer, B. J. F., Young, C. J., & Bartlett, B. J. (1989). *Memory improved: Enhance reading comprehension and memory across the life span through strategic text structure*. Hillsdale, NJ: Erlbaum.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Mislevy, R. J., Steinberg, L.S., & Almond, R.G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19,* 477–496.

Mislevy, R. J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Mochida, A., Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing, 23,* 73–98.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *Modern Language Journal, 87*, 261–276.

Nassaji, H. & Geva, E. (1999). The contribution of phonological and orthographic processing skills to adult ESL reading: Evidence from native speakers of Farsi. *Applied Psycholinguistics, 20*, 241–267.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines* (RELC supplement), *5,* 12–25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

National College English Testing Committee. (2006). CET-4 Test Syllabus and Sample Test Paper (Revised version). Shanghai: Shanghai Foreign Language Education Press.

Oberauer, K. & Suß, H. M. (2000).Working memory and inference: A comment on Jenkins, Myerson, Hale & Fry (1999). *Psychonomic Bulletin & Review 7*, 727–733.

Olson, R. K., Kliegl, R., Davison, F., & Foltz, G. (1985). Individual and developmental differences in reading disability. In G. E. MacKinnon & T. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 4, pp.1–64). San Diego, CA: Academic Press.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 174–200). New York: Cambridge University Press.

Paribakht, T. S., & Wesche, M. (1993). The relationship between reading comprehension and second language development in a comprehension-based ESL program. *TESL Canadian Journal, 11,* 9–29.

Paris, S. G., Wasik, B. A., & Turner, J. C. (1991). The development of strategic reading. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2; pp. 609–640). New York: Longman.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A. (1991). Representations and awareness in the acquisition of reading competence. In L. Rieben & C. A. Perfetti (Eds.), *Learning to read: Basic research and its implications* (pp.33–44). Hillsdale, NJ: Erlbaum.

Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp.141–180). Hillsdale, NJ: Erlbaum.

Perfetti, C. A., & Zhang, S. (1995). Very early phonological activation in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 24–33.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension sill. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp.227–247). Malden, MA: Blackwell.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(1), 26–56.

Phakiti, C. A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237–272.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8* (4), 252–266.

Pulido, D. & Hambrick, D. (2008). The virtuous circle: Modeling individual differences in L2 reading and vocabulary growth. *Reading in a Foreign Language: Special Issue on Reading and Vocabulary, 20* (2), 164–190.

Pulido, D. (2009). Vocabulary processing and acquisition through reading: Evidence for the rich getting richer. In *Second Language Reading research and Instruction: Crossing the Boundaries* (Eds., Z. Han & N. Anderson), 65–82. Ann Arbor, MI: University of Michigan Press.

Purpura, J.E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: A structural equation modelling approach. *Language Testing 15* (3) 333–379.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review 56*, 282–308.

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21* (1), 28–52.

Rayner, K., & Pollastsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10,* 355–371.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10,* 77–89.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In Theoretical issues in memory. R. J. Spiro, B. C. Bruce, & W.E. Brewer (Eds.), 33−58. Hillsdale, N. J.: Erlbaum.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (ed.), *Attention and performance VI* . Hillsdale, NJ: Erlbaum.

Sanchez, E.& Garcia, J. R. (2009). The relation of knowledge of textual integration devices t expository text comprehension under different assessment conditions. *Reading and Writing: An Interdisciplinary Journal, 29,* 1081–1108.

Schallert, D. L., & Martin, D. B. (2003). A psychological analysis of what teachers and students do in the language arts classroom. In J. Flood, D. Lapp, J. R. Squire, & J. M. Jensen (Eds.), Handbook of research on teaching the English language arts (2nd ed., pp. 31−45). New York: Macmillan.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18* (1), 55–88.

Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language specific knowledge in native and foreign language reading comprehension: An empirical study among Dutch students in Grades 6, 8, and 10. *Language Learning, 48*, 71–106.

Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Language-dependent and language-independent knowledge in native and foreign language reading comprehension: An empirical study among Dutch students in Grades6, 8, and 10. *Language Learning, 48,* 71–106.

Segalowitz, N. S. (2000). Automaticity and attentional skill in fluent performance. In H.Riggenbach (Ed.), *Perspectives on fluency* (pp.200–219). Ann Arbor: University of Michigan Press.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14,* 369–385.

Segalowitz, N. S., Poulsen, C., & Komoda, M. (1991). Lower-level components of reading skill in higher level bilinguals: Implications for reading instruction. *AILA Review, 8,* 15–30.

Share, D., & Stanovich, K. E. (1995). Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. In J.S. Carlson (Ed.), *Issues in education: Contributions from psychology* (Vol. 1; pp. 1–57). Greenwich, CT: JAI press.

Shiotsu, T. (2003). *Linguistic knowledge and processing efficiency as predictors of L2 reading ability: A component skill analysis*. Unpublished doctoral dissertation, University of Reading, United Kingdom.

Shiotsu, T. and Weir, C.J. (2007). The relative significance of syntactic knowledge and vocabulary breadth In the prediction of reading comprehension test performance. *Language Testing*, *24* (1), 99–128.

Siegel, L.S. , Share, D., & Geva, E. (1995). Evidence for superior orthographic skills in dyslexics. *Psychological Science, 6,* 250–254.

Smith, F. (1971). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. New York: Holt, Rinehart and Winston.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16,* 32–71.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp.418–452). New York: Longman.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24,* 402–433.

Syllabus for College English Test. (2006). National College English Testing Committee, Shanghai, China: Shanghai Language Education Press.

Taillefer, G. F. (1996). L2 reading ability: Further insights into the short-circuit hypothesis. *The Modern Language Journal,* 80, 461–477.

Tang, G. (1992). The effect of graphic representation of knowledge structures on ESL reading comprehension. *Studies in Second Language Acquisition, 14*, 177–195.

Taylor, B. M., & Beach, R. W. (1984). The effects of text structure instruction on middle grade students' comprehension and production of expository text. *Reading Research Quarterly, 19,* 134–146.

Torgesen, J. K., & Burgess, S. R. (1998). Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational and instructional studies. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp.161–188). Mahwah, NJ: Erlbaum.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Urquhart, A. H. & Weir, C. J. (1998). *Reading in a second language: Process, product and practice.* Harlow: Longman.

van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstjn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology, 96,* 19–30.

van Gelderen, A., Schoonen, R., Stoel, R., de Glopper, K., Hulstjn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology, 99*, 477–491.

Venezky, R. L. (1984). The history of reading research. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2; pp. 3–38). New York: Longman.

Verhoeven, L. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading, 4*, 313–330.

Vongpumivitch, V. (2004). *Measuring the knowledge of text structure in academic English as a second language (ESL)*. Unpublished PhD dissertation, University of California, Los Angeles.

Wang, C. (2006). The required conditions of foreign language learning and reflections on the reform of the College English Test. *China University Teaching (Journal in Chinese)*, 2006 (11), 48–51.

Wang, M., & Koda, K. (2007). Commonalities and differences in word identification skills among learners of English as a second language. *Language Learning, 57*: Suppl. 1, 201–222.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology, 49,* 51–79. Weir, C. J. (1983). Identifying the language needs of overseas students in tertiary education in the United Kingdom. Unpublished PhD thesis, Institute of Education, University of London.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York: Palgrave MacMillan.

Weir, C. J., Yang, H. & Jin, Y. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes. Studies in language testing 12. Cambridge:* Cambridge University Press.

Yang, H. (2000). *Validation study of the College English Test*. Paper presented at the International Association of Applied Linguistics (AILA). Waseda, Japan.

Zheng, Y. & Cheng, L. (2008). College English Test (CET) in China. *Language Testing, 25* (3), 408–417.

辜向东[Gu, X.](2005). *正面的还是负面的？大学英语四、六级考试对我国大学英语教学的反拨效应实证研究[Positive or negative? An empirical study of CET washback on college English teaching and learning in China]*.上海交通大学博士论文[Unpublished doctoral dissertation, Shanghai Jiaotong University, China].

韩宝成[Han, B.] (2002).高校学生英语能力测试改革势在必行 [The need of the reform in the English language proficiency test of Chinese undergraduates]. *外语教学与研究[Foreign Language Teaching and Research], 34* (6), 410–411.

韩宝成、戴曼纯、杨莉芳[Han, B., Dai, M., & Yang, L.] (2004). 从一项调查看大学英语考试存在的问题 [Problems of the College English Test as revealed in a survey]. *外语与外语教学[Foreign Languages and Their Teaching], 179,* 17–23.

何莲珍、张洁[He, L., & Zhang, J.] (2008). 多层面Rasch 模型下大学英语四、六级考试口语考试(CET-SET)信度研究[Investigating the reliability of CET-SET using the Multi-Facet Rasch Model]. *现代外语[Modern Foreign Languages], 31*(4), 388–398.

黄忠廉、刘丽芬[Huang, Z., & Liu, L.] (1996). 翻译：CET-4不可或缺的题型[Translation: An indispensible component of the CET]. *外语界[Foreign Language World] , 61*, 30–32.

黄忠廉、刘丽芬、倪传斌[Huang, Z., Liu, L., & Ni, C.](1996). CET-4新增翻译题型的调查与分析[Survey of the translation items of the CET-4]. *外语教学与研究[Foreign Language Teaching and Research], 107,* 61–65.

金艳[Jin, Y.] (2000). 大学英语四、六级考试口语考试对教学的反拨作用[Washback of the CET-SET on teaching]. *外语界[Foreign Language World], 80,* 56–61.

金艳、吴江[Jin, Y., & Wu, J.](1998). 以“内省”法检验CET阅读理解测试的效度[Examining the testing validity of CET reading comprehension by introspection]. *外语界[Foreign Language World] ,70,* 47–52.

刘润清[Liu, R.] (2003).高校英语教学改革笔谈之二 [Discussions about college English education reform]. *外语教学与研究[Foreign Language Teaching and Research] 35* (3), 221.

潘之欣[Pan, Z.] (2003). *交际性听力理解考试的开发-全国大学英语四、六级考试听力理解部分改革[Developing a communicative listening comprehension test: Revision of the listening comprehension part of the CET-4 and -6].* 上海交通大学博士论文[Unpublished doctoral dissertation, Shanghai Jiaotong University, China].

钱冠连[Qian, G.](2003).还是要整合性考试—谈纯分析性考试为何是失误 [In favor of integrative assessment — Exploring the causes of the failure of analytic tests]. *外语教学与研究[Foreign Language Teaching and Research],35* (5), 379−380.

王初明[Wang, C.] (2006).外语学习的必要条件与大学英语四、六级考试改革的反思[The required conditions of foreign language learning and reflections on the reform of the College English Test].*中国大学教学[China University Teaching], 2006* (11), 48−51.

王跃武[Wang, Y.] (2004). 大学英语四、六级考试作文网上阅卷试验研究[An empirical study on the CET online marking system]. *外语界[Foreign Language World], 103*, 74–79.

王跃武、朱正才、杨惠中[Wang, Y., Zhu, Z., & Yang, H.] (2006). 作文网上评分信度的多面Rasch 测量分析[Multi-facet Rasch model analysis of the scores generated by the CET online marking system]. *外语界[Foreign Language World], 111,* 69–76.

杨惠中、Cyril Weir [Yang, H., & Weir, C.](1998). *大学英语四、六级考试效度研究*［*Validation study of the National College English Test]*. 上海: 上海外语教育出版社[Shanghai, China: Shanghai Foreign Language Education Press].